Machine-Assisted Phonemic Analysis

Timothy Kempton

Department of Computer Science

February 2012

Dissertation submitted to the University of Sheffield
for the degree of Doctor of Philosophy
Supervisor: Professor Roger K. Moore

Acknowledgements

Firstly I would like to thank my supervisor Roger Moore. From the start, Roger was flexible and enthusiastic about my original proposal. I have really appreciated his patience as I slowly developed in the role of a researcher. I've particularly benefitted from Roger's wealth of knowledge of previous speech recognition work, and his depth of insight on the subject.

I am also very grateful for being funded by the UK Engineering and Physical Sciences Research Council (EPSRC grant number EP/P502748/1) through the University of Sheffield.

I have had a lot of assistance from fieldworkers in Southeast Asia. Andy Castro has been particularly helpful in providing the Kua-nsi data, and talking through the phonology of the language. Brian Crook was also on hand later to provide transcriptions and recordings on request. I have also benefited from conversations with Cathryn Yang about fieldwork and from receiving her survey data on the Nisu language.

Mary Pearce has been tremendously helpful in conversations about field linguistics and phonemic analysis, and I have appreciated her dedication in reading through an earlier draft of the thesis. Cathy Bartram gave me a helpful steer at the start on phonemic analysis heuristics. Juha Yliniemi put me in touch with the above linguists including David Morgan who helped me access the resources at the SIL UK library. I have hugely benefited from the expertise of those on the SIL linguistics discussion list. I have appreciated extended discussions with Stephen Marlett, Robert Hedinger, and Steve Parker and I am grateful for the data they sent me.

I'm thankful for the Ethnologue editorial team who ran some queries on their database for me. Brian Migliazza helped with initial queries and I value the further discussions with Paul Lewis regarding language vitality factors.

It has been a privilege to talk through areas of research with linguists in academia. Near the beginning of the PhD, I met John Goldsmith at a conference in UCL who encouraged me to pursue the original idea. Sharon Peperkamp has been very helpful in answering all my questions on her previous research. Robert Kirchner's sabbatical at Sheffield was perfect timing for me to learn phonology, and I'm grateful for the discussions we had in our shared office. Sandra Whiteside and Frank Herrmann taught me practical phonetics at Sheffield, and I appreciate their openness for further questions and discussions. I also benefited from a conversation with Ranjan Sen on phonological rules. I am also grateful to Tony Simons for sharing his past knowledge

particularly the spectrogram reading material.

Assistance for the allophones example in Chapter 1 came from Sesotho speakers Lehlohonolo Mohasi and Peter Lebiletsa. Ben Murphy recorded the northeast English at the Stadium of Light.

John-Paul Hosom, Andreas Stolcke, Petr Schwarz and Larry Hayashi have all kindly given me assistance in using their software.

Advice on evaluation metrics came from Paul Clough and Colin Champion. Colin originally introduced me to both the ROC-AUC measure and the value of statistical significance.

The SpandH Research Group has been great fun to be part of (thanks especially to James Carmichael) and a good forum for sharing ideas. I have enjoyed working with Emina Kurtic & Ahmet Aker on forced alignment, Emina has been helpful for linguistic advice and proofreading, and Ahmet providing support for text processing. Matt Gibson has provided a lot of help with anything mathematical particularly acoustic modelling. This is also true for Ning Ma who also provided insightful comments on this when proofreading. Vincent Wan gave me head start with language modelling, both in understanding it and using the correct tools. Sarah Creer was able to point me in the direction of some particularly relevant phonological research and lend me the appropriate books. I have benefited from discussions with Odette Scharenborg, particularly about the importance of comparing phoneme inventories. I also have appreciated Jan Gorisch's proofreading as well as providing an extra ear for checking my transcriptions. I am grateful to other members of this research group for productive discussions including Thomas Hain, Sue Harding, Guy Aimetti, Herman Kamper, Phil Green, Guy Brown, Robin Hofe and Jon Barker.

I am completely indebted to my family, who show me much love and patience. My brother Matthew was able to use his PhD experience to coach me, and also taught me more about statistical significance. And I'm very grateful to my mother, who did lots of typing. It's been great having regular encouragement from my father, Jessica and Joe.

It's been so good to have a supportive bunch of housemates. Tim Brown has been checking my progress and doing some of the proofreading. Philip Wilson was able to check my use of statistics and helped in finding the equation for the Hockett heuristic. I've also appreciated the moral support of Richard Wilson and Fabian Avila. It's been great to have a house linked with the church, and I'm so thankful for everyone in Broomhall gospel community past and present. I'm aware that in doing a PhD I run the risk of becoming more individualistic and self-indulgent; I am very grateful for my brothers and sisters in Christ reminded me of the gospel which keeps me sane e.g. the Richardsons had been helpful in that way over the full time period. In the wider church I've particularly appreciated the support and friendship of Lucy Mitchell, Piers & Shirley Miller, and all the Elders. Fred Hughes has been a great support long term.

As I reflect on the many people that have helped me, and the friendship of many of them, I am full of thanks to God for providing these people and giving them their gifts.

This thesis is dedicated to everyone at The Crowded House, Sheffield.

Abstract

There is a consensus between many linguists that half of all languages risk disappearing by the end of the century. Documentation is agreed to be a priority. This includes the process of phonemic analysis to discover the contrastive sounds of a language with the resulting benefits of further linguistic analysis, literacy, and access to speech technology. A machine-assisted approach to phonemic analysis has the potential to greatly speed up the process and make the analysis more objective.

Good computer tools are already available to help in a phonemic analysis, but these primarily provide search and sort database functionality, rather than automated analysis. In computational phonology there have been very few studies on the automated discovery of phonological patterns from surface level data such as narrow phonetic transcriptions or acoustics.

This thesis addresses the lack of research in this area. The key scientific question underpinning the work in this thesis is *"To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?"*. A secondary question is *"What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?"*

It is demonstrated that a machine-assisted approach can make a measurable contribution to a phonemic analysis for all the procedures investigated; phonetic similarity, phone recognition & alignment, complementary distribution, and minimal pairs. The evaluation measures introduced in this thesis allows a comprehensive quantitative comparison between these phonemic analysis procedures. Given the best available data and the machine-assisted procedures described, there is a strong indication that phonetic similarity is the most important piece of evidence in a phonemic analysis.

The tools and techniques developed in this thesis have resulted in tangible benefits to the analysis of two under-resourced languages and it is expected that many more languages will follow.

Contents

Chapter 1

Introduction

1.1 Motivation

1.1.1 Importance of phonemic analysis

Proportion of endangered languages

Throughout human history, languages have come and gone but there is a general consensus that in this century, we now face an unprecedented scale of language extinction. According to an assessment by the UN, half of all the estimated 6000 living languages risk disappearing by the turn of the century (Moseley, 2009). On average this is equivalent to one language dying out every fortnight (Crystal, 2000, p.19).

Does this matter? There has been much discussion on this subject (Crystal, 2000; Ostler et al., 2008; Grenoble and Whaley, 2006) most of the arguments can be summarised in a few points. First, language endangerment matters to the language community which is being affected. Since language forms the main cultural identity of the community, it can be very difficult for that community when it disappears. Encoded in the language is the inherited knowledge of the community to help them survive in their local environment e.g. oral traditions and vocabulary for local flora and fauna. Second, it matters to humanity as a whole. It is argued that cultural and linguistic diversity supports survival in diverse environments and therefore helps to safeguard our future as a species. Third, a loss of an undocumented language is a loss of data for theories about languages e.g. it might be difficult to construct the history of a language family when there are too many missing descendants. A recent attempt to infer a geographical region for the universal origin of language (Atkinson, 2011) depended on accurate data about phoneme inventories from hundreds of diverse languages (Dryer and Haspelmath, 2011). With less languages it would be more difficult to come to any conclusion.

One of the immediate priorities when faced with an endangered language is to document it

(Grenoble and Whaley, 2006, p.68; Crystal, 2000, p.149). The more endangered the language, the more important this is. Any further revitalisation efforts can then make use of this data. Traditionally this is in the form of descriptions such as dictionaries and grammars. In recent years, there has also been an emphasis on comprehensive documentation of language use, such as storytelling recorded on video (Himmelmann et al., 2002).

Phonemic analysis for language documentation and description

A phonemic analysis is a fundamental part of the description and documentation of a language. It sits within the broader framework of a phonological analysis which is an investigation into the whole sound system of a language. A phonemic analysis is more narrow, in that it is primarily concerned with identifying the contrastive sounds.

Two sounds contrast if substituting one for another in a word can change the meaning of that word. For example, in English the word *lip* [lɪp] has its meaning completely changed if [l] is substituted for [d]. Therefore [l] and [d] contrast; each sound is the realisation of a different phoneme; /l/ and /d/ respectively. Some sounds are articulated differently but do not contrast. For example in English the ejective [p'] i.e. produced with glottalic initiation, is occasionally used at the end of an utterance e.g. *stop* [stɒp'] (Wells, 1982, p.261), but this does not contrast with [p^h]; there is no change in meaning if either sound is substituted for the other. They are allophones; and are generally judged to be the same sound by English speakers. They are both realizations of the same phoneme, /p/.

Sesotho, a language spoken in Lesotho, has similar sounds but they contrast differently. In Sesotho [l] and [d] are allophones but there is a contrast between the sounds [p^h] and [p'] (Demuth, 2007). This is shown in Figure 1.1 with example words in Table 1.1. No previous illustrations could be found in the literature showing cross language phonemic effects both ways with real words, so this example[1] was compiled with the assistance of indigenous speakers from Lesotho and Northeast England (Sunderland).

The process of a phonemic analysis is described more fully in Section 1.2. A phonemic analysis leads to at least three important follow-on benefits for a language; further linguistic analysis, literacy, and speech technology (Figure 1.2).

[1]Northeast English is used in this example because the accent shows a very similar vowel to Sesotho; /o/ rather than /əʊ/ as in RP English (Wells, 1982; Watt and Allen, 2003). The Sesotho name Polo is short for the full name Polomakhoashe (written in the Lesotho orthography rather than the South African variant). The utterance [bolo] which is a nickname in English is included to confirm that there is a three way contrast for bilabial plosives in Sesotho but only a two way contrast in English. Bolo not a common English name but at the time of writing it is the nickname given to Boudewijn Zenden, a Dutch football player at Sunderland AFC.

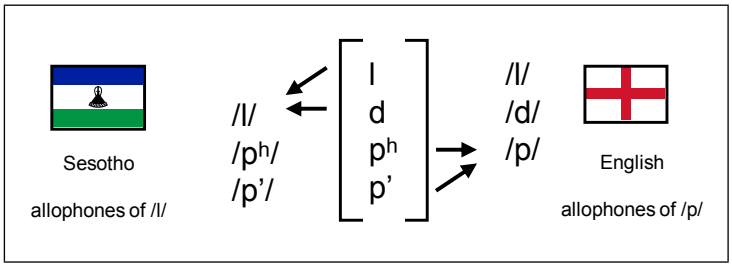

Figure 1.1: Sesotho and English allophones

Utterance	Sesotho interpretation		English (NE) interpretation	
[li]	/li/	them (obj. concord)	/li/	lea (meadow)
[di]	/li/	them (obj. concord)	/di/	Dee (UK river)
[pʰolo]	/pʰolo/	ox	/polo/	polo (sport/mint)
[p'olo]	/p'olo/	Polo (name)	/polo/	polo (sport/mint)
[bolo]	/bolo/	ball	/bolo/	Bolo (name)

Table 1.1: Sesotho and English perceptions of the same utterance

Follow-on benefits: further linguistic analysis

A phonemic analysis forms an initial understanding of the phonology of a language and lays the groundwork for further language description (Hayes, 2009, Ch.2; Gleason, 1961, Ch.1,2,17). This could include more phonology such as detailed analysis of the suprasegmentals; sound patterns that span longer time sequences than phones (such as stress and intonation). A good understanding of phonology can also help with the analysis of word components i.e. morphology and the practical task of making dictionaries. In turn, this can lead on to syntactic (sentence structure) and semantic (sentence meaning) analysis. Knowledge of the phonology is also necessary for a detailed phonetic analysis of the language (Ladefoged, 2003, p.1). Historical linguistics depends on a good understanding of sound change, so a knowledge of the contrastive sounds system of each language is invaluable for this type of research (Arlotto, 1981).

Follow-on benefits: literacy

Since phonemic analysis can uncover the set of contrastive sounds in a language i.e. the phoneme inventory; this process can be used to construct an alphabet. In the past this was the principal use of a phonemic analysis (Pike, 1947). Even now, in modern times a phonemic analysis is currently the most flexible and efficient method to establish a writing system (Werner, 2000, p.62; Hayes, 2009, p.47). Of course many other factors are brought into play in developing an orthography, such as morphology, sociolinguistics and government policy but a phonemic analysis lays the theoretical groundwork. There is still a great need for writing systems since

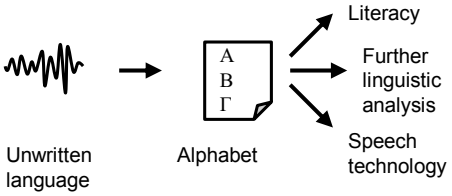

Figure 1.2: Phonemic analysis enables other important developments

only 42% of all languages are known to have them (Lewis, 2009)[2]. It can be argued that some languages are better represented with a non-alphabetic writing system e.g. a syllabary or logosyllabary (Daniels and Bright, 1996, p.4) but even in this situation, conducting a phonemic analysis will help to inform this decision.

Follow-on benefits: speech technology

Without a writing system, most speech-recognition technologies are of little use i.e. speech-to-text and text-to-speech is meaningless if there is no text. And, as argued above, if text is needed, a phonemic analysis is needed. Even without a need for text most of the speech recognition tasks will have a requirement of some underlying symbolic representation which, like text will presuppose a phonemic analysis. A phonemic analysis also has the potential to improve speech recognition performance on languages that already have writing systems. For example some accents of English have slightly different phoneme inventories when compared to the inventory of a so-called standard accent commonly used in a speech recogniser. If important contrasts are not reflected in the underlying phoneme inventory then traditional modelling and adaptation techniques (e.g. alternative dictionary pronunciations, speaker adaptation) will always be suboptimal (Huckvale, 2004). For example a speech recogniser such as CMU Sphinx based on US English with a 39 phoneme inventory cannot fully model the larger inventory for RP English. The solution is to use the phoneme inventory of the target accent. For many accents, this may not be well documented, and a phonemic analysis is needed. This is also true for speech synthesis; knowledge of the phoneme inventory and associated allophonic rules are vital for modelling or adapting the lexicon, although documentation is often lacking (Fitt and Isard, 1999).

Even well documented accents need to be re-analysed at some stage because of sound change. One of the differences between most US accents and RP English is due to a number of changes in the RP accent during the 1700s which culminated in R-dropping (Wells, 1982, p.218). /ɹ/ was lost before consonants and word boundaries. This in turn ended up creating some new vowels in the RP accent. For example, the pronunciation of the word *beard* changed: /biɹd/ → /bɪəd/ and the diphthong /ɪə/ became a new phoneme. Wells (1982, p.259) states that a similar

[2]Personal communication (2010) with the Ethnologue editorial team who conducted a database search to confirm this figure

development in London English with L-vocalisation has the potential to change the future vowel system again. For example the pronunciation of the word *milk* appears to be changing: /mɪlk/ → /mɪʊk/ and the diphthong /ɪʊ/ could become a new phoneme. A phonemic analysis could be used to detect and characterise such developments.

1.1.2 Importance of *machine-assisted* phonemic analysis

Speeding up routine and tedious tasks

The process of a phonemic analysis involves looking for evidence of contrast between every possible pair of sounds. Although there are short cuts, the full analysis is a lengthy and tedious process (Hayes, 2009, p.40) which would benefit from some automation. The length of time a phonemic analysis takes is difficult to quantify because it depends on a number of factors. Hockett (1955) estimated that it takes an experienced linguist about 10 days of hard work to complete 90% of an analysis, an additional 100 days to complete 99% of the analysis and sometimes years to achieve 100%. 10 days is also a figure referred to by Pike who describes it as the length of time for trainee linguists to develop a basic albeit incomplete analysis (Pike, 1947, p.ix). Hayes writes that a full analysis can take years (Hayes, 2009, p.34) often because the linguist fails to notice a rare or difficult-to-hear contrast. Contemporary field linguists[3] confirm that such failures can lead to large scale revisions of the phonology; making time estimations difficult. However, there does seem to be some consensus about the 10 day figure for a 90% analysis, not including data collection and interaction with native speakers (which could take up to an additional 10 days). The same field linguists report that languages with particularly complex phonologies can take much longer.

There are tools to help speed up the process; such as Phonology Assistant (SIL, 2008) which provides search and sort database functionality specifically for the task of phonemic analysis. It is acknowledged as a useful tool (Dingemanse, 2008). However, it doesn't perform any automated analysis which could further speed up the routine and tedious tasks.

Greater consistency on acoustics and analysis

Cross-language bias is another issue that can affect a linguist's phonemic analysis. It is possible this could be improved with a machine-assisted approach. Each linguist will have a bias towards their mother-tongue or other languages they have experience in, when interpreting the acoustic data. This is particularly the case with difficult-to-hear contrasts. For example Hayes (2009, p.48) has described the near impossibility for himself as an English speaker to distinguish dental stops and alveolar stops which are contrastive in a dialect of Bengali but not in English. Every phonetic transcription will be effected by the bias of the linguist who wrote it. There can also

[3]This section was informed by correspondence with field linguists from SIL International

be a bias in the other parts of the analysis. For example, related languages can be very useful because the phonologies are often similar, but there is a danger that the linguist overestimates this effect and takes short cuts in the analysis that are not warranted, giving incorrect results. It is hard to predict the effect of cross-language bias, and there are rarely the resources to perform multiple independent analyses. Of course, a machine based approach may also have certain biases, but these are more likely to be consistent and repeatable. There is also the scope to combine multiple machine based approaches to reduce bias.

1.2 What is involved in a phonemic analysis?

In looking to automate phonemic analysis, it is helpful to understand the process in more detail. The process is summarised in Figure 1.3.

The phonetic stage

One of the first stages in a phonemic analysis is to take an impressionistic phonetic transcription of the language. Initially this is elicited from a text or wordlist in a language that is common to both the linguist and language consultant (i.e. the indigenous speaker). The finished wordlist would contain the word in the trade language and a phonetic representation of the target language.The wordlist is carefully chosen to reduce the chance of including loan words e.g. words such as *dog, louse, tree* (Swadesh, 1955) are used rather than modern words such as *computer*. Loan words could introduce phonological patterns that are not fundamental to the language. It is important to capture as much detail of the sounds as possible, since it is not known beforehand which sounds are contrastive (Gleason, 1961). For example, if there was no prior information about English (or Sesotho) phonology all the sounds such as [l,d,p^h,p̚] would need to be carefully transcribed. This is usually done by an experienced phonetician, who tries to be objective in minimising phonological bias from their knowledge of other languages. As stated earlier this can be a challenge, especially when attempting to detect possible contrasts not in the phonetician's language (Pike, 1947, p.67).

When identifying a sound sequence, the appropriate number of phone segments will also be identified. This segmentation can be ambiguous. For example, the same utterance of the word *year* in RP English might be transcribed as [jɹə] or [jə]. This ambiguity can often be clarified in the subsequent stage of analysis.

It is not always clear where a phone starts and ends but once this has been decided, phonetic transcription can be aligned with the audio data. This procedure of alignment is optional but it can be helpful for acoustic analysis such as vowel formant plots (Ladefoged, 2003, p.192).

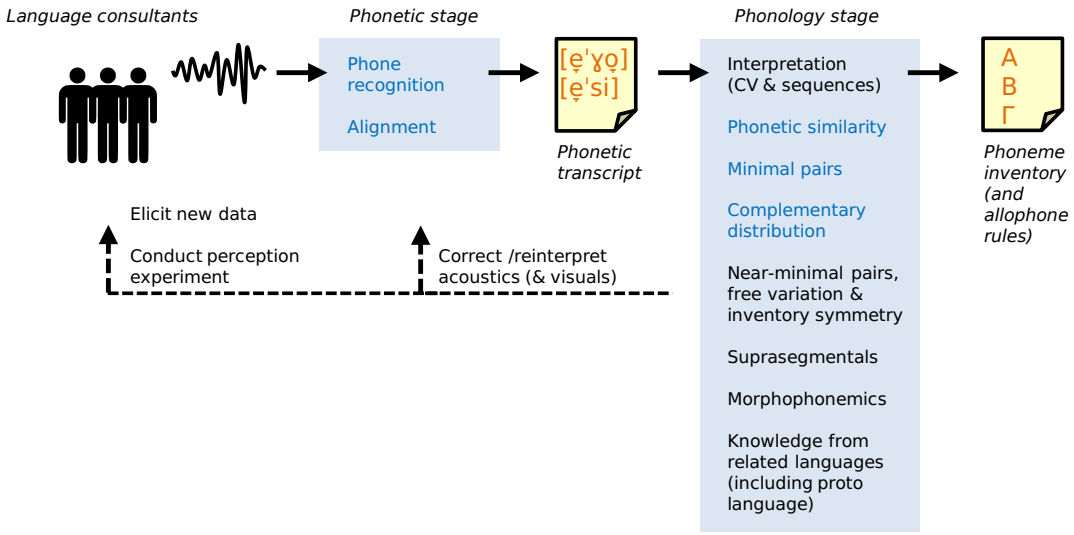

Figure 1.3: The stages in a phonemic analysis. Procedures written in blue (or grey if in monochrome) are those investigated in the thesis.

The phonology stage

Once a detailed phonetic transcript has been attempted, the analysis is primarily phonological. It is helpful to first identify ambiguous sounds (Burquest, 2006, p.164; Bartram et al., 2008). A sound may be ambiguous because it is unclear if the sound is behaving as a vowel or consonant. Or a sound may be ambiguous because it is unclear if the sound is behaving as a single phone or a sequence of phones. As in the example above with the word *year*, there may also be ambiguity as to whether a phone exists or is merely the manifestation of a transition from one phone to another. There are a number of lines of evidence that can help to indicate the best interpretation. A common approach is to investigate syllable structure (Burquest, 2006, p.155). Some conclusions of the syllable structure can be reached by analysing unambiguous phones which can in turn help with ambiguous phones. For example in Kua-nsi, a Tibeto-Burman language (Castro et al., 2010) the analysis of unambiguous phones such as open vowels and stop consonants indicates that syllables are constrained to a simple structure without consonant clusters. However, there is an apical voiced fricative [z̩^{j}] following some consonants such as [s] that appears to violate this constraint. But there is no violation if this ambiguous fricative is treated as a vowel rather than a consonant. In fact this is so common in Sino-Tibetan languages that sinologists use a dedicated (non-IPA) symbol for this vowel [ɿ]. In Kua-nsi a vowel interpretation is confirmed by acoustic evidence i.e. the sound behaves as a syllable nucleus and is consistently tone-bearing. The word *blood* [sɿ^{21}] is an example (where the superscript numbers indicate a

low and falling pitch (Chao, 1930)).

The syllable structure of Kua-nsi also helps to clarify ambiguities regarding possible sequences of phones. The simple syllable structure is apparently violated by the pair [pf]. However there is no violation if it is treated as a single consonant affricate [p͡f]. Sometimes the interpretation is a choice between equals. When this is the case, whatever choice is made, it is important to be consistent and further stages of analysis can clarify if the decision was correct[4].

After deciding on an initial interpretation of ambiguous sounds, a comparison of every sound can be made. Strictly, every phone needs to be compared against every other phone to determine whether they are phonemically distinct or not. However, in practice sounds that are phonetically very distant from each other are assumed to be phonemically distinct e.g. [t] and [m]. Relying on some notion of phonetic similarity is sometimes implicit in a phonemic analysis, but it is always important (Pike, 1947, p.69; Burquest, 2006, Ch.2; Hayes, 2009, p.54).

The principal method of determining a contrast between sounds is to find minimal pairs. These are pairs of different words that only differ by a single phone. Finding such words establishes that the phonetic difference between the two phones is contrastive. For example, consider the two English words:

[sɪp] sip
[ʃɪp] ship

These two words establish that the phones [s] and [ʃ] contrast with each other. However, it is important to look for more than one minimal pair.

Phonetically close sounds that cannot be shown to contrast using the minimal pair method could be allophones. For example, in Sesotho it is not possible to find minimal pairs that show a contrast between [d] and [l]. Their status as allophones can be confirmed if they can be shown to be in complementary distribution, meaning they appear in mutually exclusive phonetic environments. Testing for this involves listing environments for each phone i.e. the preceding and succeeding sounds. When this is done on Sesotho it becomes clear that [d] only occurs before high vowels, and [l] occurs everywhere else. This complementary distribution confirms that the two sounds do not contrast, and instead there is an allophonic relationship between them; they are both realisations of the /l/ phoneme.

At this stage, if there is still uncertainty, other less definitive analysis procedures can be used. This includes near-minimal pairs, checking for free variation, and looking for inventory symmetry. More information on these procedures can be found in Burquest (2006) and Hayes (2009).

An initial investigation of suprasegmentals is needed because they can play a part in the identification of phonemes. For example, vowel harmony in Chadic is a prosodic process, but it is a factor in deciding how many vowels there are. Likewise tone and voicing interaction can

[4]Pearce, M. (2012), personal communication

affect the number of phonemes.[5] When investigating suprasegmentals, the linguist should also be aware of autosegmental phonology. This is a phenomena where certain features such as tone or nasality can be viewed as acting independently from particular phones.

If there is some existing knowledge of the morphology of the language then the use of morphophonemics can also help. Morphophonemics cover the important interactions between morphology and phonology. An understanding of this interaction can help in the process of phonemic analysis, particularly in the area of phonological alternation and for suggesting diagnostic wordforms to elicit new data from a speaker. This diagnostic approach allows a fairly precise rearrangement of phoneme sequences which can help determine how particular phonemes behave in specific environments (Hayes, 2009, p.123).

Related languages often have similar phonologies (see for example Castro and Chaowen, 2010); so they can be helpful for suggesting hypotheses about the phonology that can be tested, rather than starting from scratch. Sometimes a historical analysis will have been conducted in the language family; giving a hypothesised proto form. Although care should be taken with such a hypothesis, a proto language can be very useful because the phonology under study could be related to the proto language via some simple transformations.

This phonology stage of a phonemic analysis is an iterative one. For example, it's possible that mistakes will be made in the interpretation stage that will only be made clear later in the analysis. When this happens the linguist will go back and try an alternative interpretation.

There can also be iteration in the wider process and this is shown in Figure 1.3 as dashed lines. Sometimes there needs to be a correction to a transcription or a reinterpretation of the original acoustic recording (or video). Sometimes further work with the language consultants is needed e.g. to conduct a perception experiment, or to elicit new data. This interactive process could also include informal conversation with the speakers.

1.3 Can a machine-assisted approach help?

The above background information on phonemic analysis leads to the following scientific question:

> *"To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?"*

The analytical procedures investigated and evaluated in this study are:

- Phonetic similarity
- Phone recognition and alignment

[5] Pearce, M. (2012), personal communication

- Complementary distribution
- Minimal pairs

To answer this question, a suitable evaluation metric for accuracy is needed. This is introduced in Chapter 3. Time savings also need to be considered and this is discussed in Chapter 7. In evaluating these individual procedures a secondary question emerges:

> *"What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?"*

1.4 Scope of thesis

The structure of the thesis is driven by the scientific question above. The subsequent chapter is an investigation into previous related work and further chapters are devoted to each of the above four procedures. A final series of discussions and conclusions form the last chapter.

The title of this thesis is "Machine-Assisted Phonemic Analysis"; the overall aim is not to fully automate the analysis, but to provide a useful tool to the linguist. There are other procedures in a phonemic analysis that are not investigated in the thesis such as those written in Figure 1.3 that are not highlighted. The four procedures stated above were chosen because they are the most mechanical and tedious procedures to perform manually, and would benefit the most from becoming partly automated.

Even restricting the scope to these four procedures, there is much uncharted territory and the emphasis in this thesis is to cover a lot of ground, sometimes at the expense of depth because it was decided that this was the best way to contribute to a field that has received little attention in the past.

The techniques developed in this thesis are relevant to all speech sounds, but parts of the evaluation focus on consonants (e.g. especially Chapters 3, 5 and 6). This is because, for the vowel data, there is some variability or uncertainty of vowel ground truth labels in the best corpora currently available. For example the TIMIT corpora of US English (Garofolo et al., 1993) covers a range of dialects and idiolects, some exhibiting the caught/cot merger, and some not (Labov et al., 2006). Also, in the Kua-nsi corpora (Castro et al., 2010; Lehonkoski et al., 2010) there is currently some uncertainty regarding the phonology of the high back unrounded vowels in the dialect used for the phonetic transcription. There is much more certainty about the phonology of the consonants. As more structured data becomes available in the future, vowels can be similarly evaluated. In the meantime it should be noted that the mean ratio of consonants to vowels over all languages is estimated at 4.25 (Dryer and Haspelmath, 2011, Ch.3), thus the lack of experiments on vowels in parts of the evaluation should not be regarded as a major problem.

A phonemic analysis is needed in at least two different practical scenarios. In finalising a writing system for language, it is important that the phonemic analysis is as accurate as possible. In this scenario, interaction with mother-tongue language speakers will be extensive and last over a prolonged period of time. A different scenario exists when conducting a survey of a number of dialects or languages, it can still be helpful to provide a rough sketch of the phonology. This must be done with less interaction from the speakers; because there is often only time for a single short visit to each village. The approaches developed in this thesis can be used in both scenarios but they are most relevant to the latter one. In this survey-scenario a single pre-defined word list is provided, and the task is to conclude as much about that phonology as possible, before further interaction with mother-tongue speakers.

The phonological framework used in this study is very much affected by the practicalities of developing a writing system for a language. The phonology is literacy and alphabet-orientated; and therefore segmental. The focus of the machine-assisted analysis is on processes that are phonologically very close to surface forms. The output is expressed as allophonic relationships between sounds rather than multiple levels of rules. This allows the linguist to finish the analysis in whatever phonological framework is appropriate whether that is rule-based (e.g. classic generative phonology) or constraint-based (e.g. optimality theory).

The emphasis on endangered languages should not detract from the much wider area of application of this work. Endangered and under-documented languages present an interesting problem where there is very little linguistic information available; often only the speech itself. Tackling such a problem requires engagement with the fundamentals of spoken language without making language specific assumptions common to most speech recognition research and some computational phonology research. This means a phonemic analysis is relevant to all languages and a machine-assisted approach to phonemic analysis could have a wide impact. So although the main application of this work is in language conservation, there is also a fresh perspective on areas of application for common languages such as English and its variety of accents.

1.5 Definitions of key phonological terms

For the purpose of clarity, some of the key phonological terms used in this thesis are defined below. There can be some variation in the literature so the principle here is to follow the definitions of Hayes (2009) where they are available, and to use other sources where it is believed they can form part of a compatible framework.

- Phoneme – a basic speech sound, a minimal unit serving to distinguish words from each other; an abstract phonological category (Hayes, 2009, p.20,p.23).

- Allophone – variant of a particular phoneme; a concrete observable sound (Hayes, 2009, p.23).
- Phone – a speech sound; the smallest discrete segment of sound in a stream of speech (Oxford, 2010).
- Free variant (as in free variation) – allophone unconditioned by its phonetic environment, i.e. freely fluctuating (Clark et al., 2007, p.116; c.f. Hayes, 2009, p.59).
- Contrastive – used to differentiate between different morphemes (Gussenhoven and Jacobs, 2005, p.49).
- Phonemically distinct – belonging to different phonemes (e.g. in English [h] and [ŋ] do not contrast but they do belong to different phonemes (Hayes, 2009, p.54)).

The last definition has been made more specific for the purpose of this thesis. Other authors may possibly view the last term as being equivalent to the term contrastive (Hayes, 2009, p.20).

1.6 Chapter summary

Half of all languages risk disappearing by the turn of the century. The process of a phonemic analysis can help in the documentation of these languages by describing the contrastive sounds. Benefits to the language (whether endangered or non endangered) includes further linguistic analysis, literacy, and speech technology. A machine-assisted approach to phonemic analysis has the potential to greatly speed up the process and make the acoustic analysis more objective.

The scientific question of this thesis is “To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?” and the procedures investigated in this study are highlighted in Figure 1.3. A secondary question is “What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?”.

The scope of this study is focused on assisting a linguist in completing a phonemic analysis rather than automating it. The primary scenario for using this tool is expected to be in a survey to obtain a rough sketch of the phonology. The practicalities of the scenario mean that the phonological framework employed is primarily pragmatic in developing a writing system but is flexible enough for the linguist to complete the analysis in whatever theoretical framework is appropriate. Although the emphasis in this thesis is on endangered languages, the principles and practical methods for deriving a phonemic analysis apply to all languages.

Chapter 2

Related work

2.1 Literacy and endangered languages

Languages can be endangered, under-documented, or unwritten. It is helpful to know if and how these factors are related; especially since it has been proposed that a phonemic analysis has a bearing on the latter two issues.

A UNESCO report (Brenzinger et al., 2003) identifies nine key factors in determining language vitality. These are shown in Table 2.1. In the report, the authors are careful to state that no single factor should be taken on its own in assessing language vitality. However they acknowledge that the first factor intergenerational language transmission (i.e. parents transmitting the language to their children) is the most common one for making an assessment, and the first six factors are especially useful for assessing language vitality. The other factors are less so; factors seven and eight are for assessing language attitudes, and factor nine is for assessing the urgency for documentation. This grading of factors, is shown in the second column of the table. Grenoble and Whaley (2006) also make an assessment which is shown in the last column.

Grenoble and Whaley agree with the UNESCO assessment that intergenerational transmission is the strongest factor. There is agreement that documentation *per se* is not a strong factor, and both studies take the view that the availability of literacy materials is strongly related to language vitality.

There has also been a study evaluating the UNESCO framework by investigating its use on 100 languages, (Lewis, 2006). It was found that, although definitions could be clarified, generally the vitality factors were suitable for characterising endangered languages. The publication by Lewis (2006) includes the raw data for the 100 languages. Although not an intended goal of the original publication, it gives an opportunity to investigate the correlation between UNESCO factors. Given the sampling criteria that Lewis uses, (no large international languages, 20 languages per continent, countries with high language diversity and several languages which have

	Factor	UNESCO	G & W
1	Intergenerational language transmission	strongest	strongest
2	Absolute number of speakers	strong	weak
3	Proportion of speakers within the total population	strong	strong
4	Shifts in domains of language use	strong	strong
5	Response to new domains and media	strong	strong
6	Availability of materials for language education and literacy	strong	strong
7	Governmental and institutional language attitudes and policies including official status	weak	strong
8	Community members' attitudes toward their own language	weak	strong
9	Amount and quality of documentation	weak	weak

Table 2.1: UNESCO Language vitality factors including assessment by Grenoble and Whaley

been subject to revitalization efforts); an attempt can be made to reach some conclusions about the relationships between the factors.

The analysis for the work of this thesis began by addressing the sparse nature of the data. For each of the 100 languages in the original study, there is not always data for every factor. *Little's MCAR test* (Little, 1988) was used as the appropriate test for assessing the impact of the missing data. The test showed that the null hypothesis, i.e. that the data was missing completely at random, could not be rejected. This meant a further analysis of correlation could proceed. Since the factors used ranked values, an analysis of correlation between factors was conducted with *Spearman's rho*, using pairwise deletion for missing values. This gave a number of apparent correlations but, because of the issue of multiple comparisons, some of these may occur by chance. To allow the reader to test a hypothesis, a single pair of factors should be chosen to check for correlation before viewing the results in Figure 2.1 (only statistically significant correlations are shown i.e. $p < 0.05$). The chart is similar to the triangular mileage charts on street atlases which have the cities listed on the diagonal. Here the UNESCO factors are listed in place of cities, and the correlations listed in place of the mileage.

Given that intergenerational transmission was already judged to be the strongest factor of language vitality (Table 2.1), and that the factor of particular interest is literacy; it is legitimate to test for a single correlation between factor 1 and 6. Figure 2.1 shows there is a small but significant correlation between these two factors ($\rho = 0.36$, $p < 0.05$) . Note that factor 6 is not just about materials but does include literacy *per se* "[a score of 4 means] ... children are developing literacy in the language" (Brenzinger et al., 2003).

For a more exploratory data analysis, Figure 2.2 highlights existing correlations after the Bonferroni correction for multiple comparisons. Two clusters are shown; with literacy related to

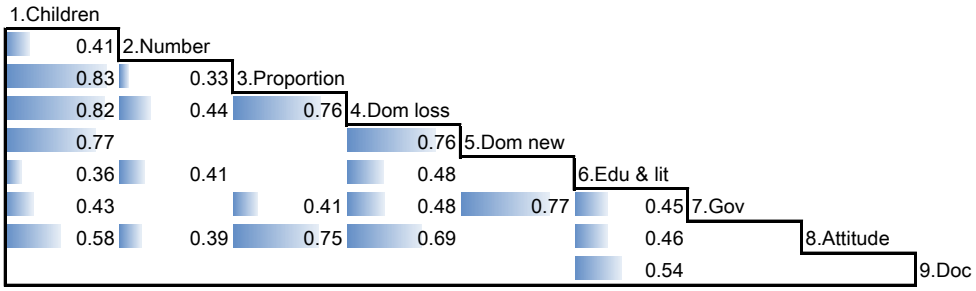

Figure 2.1: Correlations of the UNESCO factors based on Lewis's data indicates a link between literacy materials and intergenerational transmission of language (only single previously decided comparisons can be made). See Table 2.1 for a fuller description of the factors.

documentation only. Without any prior decision to test for a specific correlation, only the factors within clusters can be said to be correlated. If there was a need to give a similar assessment of factor strength as in Table 2.1, given the assumption that intergenerational transmission (factor 1) is the strongest, the correlations suggest that only the proportion of speakers, domain loss, and community attitudes (factors 3,4,8) are also strong factors.

This statistical analysis of correlation is an important contribution in assessing the relative importance of the UNESCO language vitality factors. The results have a bearing on the link between literacy and vitality which is relevant to this thesis, but it is also expected that the findings reported here will be of significance in the broader area of language conservation.

A recent proposal for a new evaluative framework for language endangerment (Lewis and Simons, 2010), suggests that literacy is an important factor for vitality but only when there is a sufficient level of intergenerational transmission already. Modeling such a dependency may provide evidence of stronger correlation but there was not enough data in the 100 language points to come to a conclusion either way. More surveys are needed. However, for the moment there does seem to be a consensus (and some statistical evidence) that literacy is a factor in language vitality.

It is difficult to come to a reliable figure of how many languages are unwritten. As noted in Chapter 1, a search of the Ethnologue database (Lewis, 2009)[1] reveals that 42% of languages are recorded to have at least one writing system. This leaves 58% that are either unwritten or lacking information on their literacy status. Therefore there is arguably a great need for any tool that will help develop writing systems.

[1]Personal communication (2010) with the Ethnologue editorial team who conducted a database search to confirm this figure

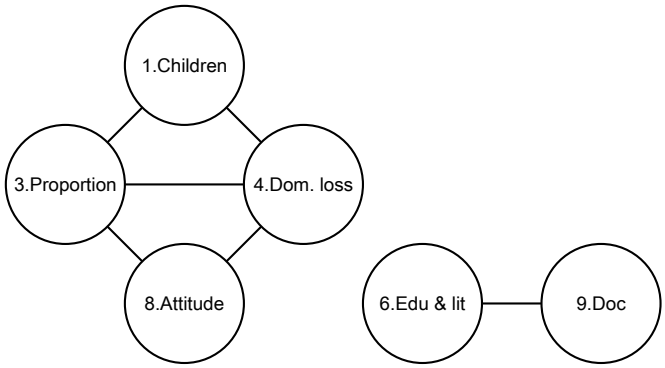

Figure 2.2: Correlated UNESCO factors based on Lewis's data after corrections for multiple comparisons. See Table 2.1 for a fuller description of the factors.

2.2 Phonemic analysis

The practice of phonemic analysis has its roots in the American structuralist tradition (Sapir, 1925). Pike (1947) outlines the process in detail in a book aimed at field linguists developing writing systems. It contains many pragmatic heuristics and includes a number of exercises and drills, although most are based on a hypothetical language. A briefer than more formal account from the structuralist school is given by Gleason (1961) (cited in Hayes (2009)).

The concept of the phoneme has been controversial in phonology, and therefore the process of phonemic analysis has also been questioned. In the early years, phonemic analysis was developed in the framework of taxonomic phonemics; which stressed a taxonomy of three levels that should be kept separate; phonetics, phonemics, and morphophonemics (Odden, 2005). Transformations between levels had different mathematical rules, such as the principle of biuniqueness which was enforced between the phonemic and phonetic level. This meant that an allophone could only belong to one phoneme. Biuniqueness was not enforced between the morphophonemics and phonemics boundary which allowed neutralization at this higher level (e.g. word-final devoicing in German). There was a growing optimism of what could be achieved with this formalised method. Phonemic analysis was seen as a simple mechanical process which potentially could be completely automated, for example without any reference to minimal pairs (Bloch, 1948), and with an almost exclusive use of complementary distribution (Harris, 1951). Halle (1959) showed that keeping a strict separation between morphophonemics and phonemics didn't work; it produced overly complicated and unintuitive rules for voicing assimilation in Russian. In giving a comprehensive account of these historical developments, Odden (2005) points to a similar problem with US English flapping. A strict separation between levels is not tenable because it predicts the nonsensical interpretation that there is a separate flap phoneme in US English. Without a strict separation between levels of analyses, biuniqueness could no

longer be held on a single level as a rule for allophones. Alongside other ambiguities this resulted in a less deterministic procedure for phonemic analysis. With the abandonment of taxonomic phonemics there was now less optimism about automation.

Despite the change of emphasis, the contemporary procedure of a phonemic analysis (Burquest, 2006; Hayes, 2009) is still very similar to earlier work. There is still a structured procedure for the linguist to follow, but rather than steps of self-contained analysis, there is an emphasis on iterative cycles spanning the whole process and bringing in many knowledge sources as described in Chapter 1.

2.3 Directly relevant work

2.3.1 Software to help with phonemic analysis

There has long been a recognised need for computer tools to assist with phonemic analysis. From conversations with a number of field linguists, and a search through the literature it appears that the first tool available to help in a phonemic analysis was *Findphone* developed at SIL in the UK during 1984 (Hunt, 2008). At the heart of this tool was a powerful search feature to find any combination of transcribed phones in any environment. After the final MS-DOS version in the mid-90s (Bevan, 1995), a number of attempts were made to fill the void on other platforms. However, linguists valued the functionality of Findphone to such an extent that some were still using it a decade later. It was at this point that a suitable successor for the Windows platform had been developed which met the functionality (Hunt, 2008). This is called *Phonology Assistant* and is now Unicode compliant (SIL, 2008). Given a phonetically transcribed word list, Phonology Assistant provides a range of database tools to help with the analysis. A phone inventory is automatically derived and can be displayed as consonant and vowel charts with histograms. Regular-expression-like searches that include articulatory features, allow environment charts to be quickly explored. There is also functionality to help with identifying minimal pairs. All the analysis is on transcripts, but if there are audio recordings associated with the word list, these can be played. The interface allows all this functionality to be linked together in an intuitive way which has been particularly appreciated by linguists (Dingemanse, 2008).

Currently there are two other similar tools; Dekereke (Casali, 2009) which has a particular strength in investigating phonotactic generalisations, and PTEST (Phonology Template Editor and Search Tool) (SIL, 2010), which has the ability to search for a number of predefined phonological rules and tabulate the results in a report. Both these tools have functionality that is targeted at African languages.

There are a number of tools to assist the linguist in phonetics especially in the area of detailed acoustic analysis. A notable example is Praat (Boersma and Weenink, 2011), a powerful tool with a long history of usage by linguists.

All this software certainly helps speed up the work of a phonemic analysis. The functionality is particularly tailored to phonetic data, but it is essentially database functionality such as search and sort that is being offered. The tools stop short of doing the analysis themselves. It is possible that tools making use of computational phonology could help the linguist further by performing part of the analysis automatically.

2.3.2 Computational phonological analysis

Much current work on computational phonology has its focus close to the phonology-morphology boundary e.g. as evidenced by the majority of publications from the Association for Computational Linguistics (ACL) special interest group on computational morphology and phonology (SIGMORPHON). The few experiments closer to the phonology-phonetic boundary have been carefully supervised i.e. the learning algorithm having knowledge of both the underlying and surface forms whether the learning algorithm is a traditional transducer (Gildea and Jurafsky, 1996) or is in the optimality theory framework (Boersma et al., 2003).

Peperkamp et al. (2006) investigated the problem of discovering allophones and their associated rules without knowledge of underlying forms: a much more unsupervised process of learning. The study was conducted in the context of modelling infant language development. It is also particularly relevant to a phonemic analysis where the linguist does not know a priori what the underlying forms are. One limitation of this particular study, was that the phonetic data was synthetically derived from a phonemic transcription in the first place. As in the other two learning studies (Gildea and Jurafsky, 1996; Boersma et al., 2003) it was found that the general learning algorithm benefited from linguistically motivated biases in the learning process.

There has been a small amount of work looking at automatic speech recognition assisted computational phonology. For example Tajchman et al. (1995) calculated the probability of the occurrence of pre-defined phonological rules on a speech corpus and Lin (2005) investigated the link between speech recognition and infant speech perception using acoustically derived sub-word units (ASWUs). Interestingly at the end of his PhD thesis Lin states that “allophones, [and other phonological phenomena]... may eventually find their place in the model”. Related to this, exemplar theory (Bybee, 2001) has a greater emphasis on surface forms. Kirchner et al. (2010) describes experiments on pattern entrenchment; e.g. where a cluster of exemplars for a particular word already sharing a bias for a phonological pattern becomes more biased over time. These exemplars were whole words of actual acoustic data from a single speaker. It is hoped that in future, their algorithm will work across all input data without the artificial supervision of keeping clusters separate for each word. The *emergent structure* described by Kirchner et al. (2010) can be viewed as ASWUs (see Section 2.4.1).

Within this field of computational phonology, the work of Peperkamp et al. (2006) is most

relevant to the problem of phonemic analysis. The above studies incorporating acoustics are also helpful because they indicate that it is possible to integrate such data into the model. However, if a system is going to be built that can robustly handle large amounts of diverse acoustic data (rather than investigating highly supervised one-off experiments), it is helpful to look at the associated field of automatic speech recognition.

2.4 Speech recognition technology

2.4.1 Multilingual acoustic modelling for ASR

Automatic speech recognition (ASR) especially large vocabulary continuous speech recognition (LVCSR) most commonly uses phoneme-orientated acoustic models (Ostendorf, 1999). These models are typically trained with acoustic data alongside sequences of phoneme labels derived from the pronunciation dictionary. A model trained for each phoneme is called a context independent or monophone model. If multiple models are built for each phoneme depending on the neighbouring phonemes then the models are called context dependent. Context dependent acoustic models such as tied state triphone models are generally more accurate than monophones because they model the contextually conditioned variation (Jurafsky and Martin, 2008). This basic approach to acoustic modelling has been shown to work well across different languages by a number of research groups (Schultz and Kirchhoff, 2006, p.76). The acoustic features are usually MFCCs (mel frequency cepstral coefficients), although for tone languages, pitch information is often also added (Lei et al., 2009).

The success of sharing a common architecture across languages has prompted some to investigate whether models can be shared across languages as well (Schultz and Kirchhoff, 2006, p.77). There is an inherent difficulty with direct sharing of phoneme-based models across languages, because a phoneme is specific to a language. This is why phoneme-based models do not transfer well between languages (Williams et al., 1998). When Schultz and Waibel (1998) combined phoneme-based models in an unsupervised manner into a decision tree they found that many of the initial questions in the tree were about the language, confirming the language specific nature of a phoneme.

Phone-based modelling

There have also been studies on building phone-based models for multiple languages, and some phoneme-based attempts probably fit in this category. This approach uses similar sounds or phones that are shared between languages. These range from specific multilingual models which are aimed at model sharing between a closed set of known languages (Burget et al., 2010), to cross-language transfer where there is no labelled training data in the target language. Cross-language transfer has been attempted with a single source language (Lööf et al., 2009), where

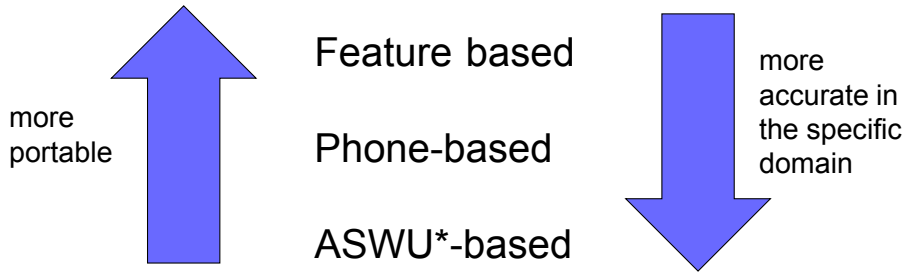

Figure 2.3: Multilingual Acoustic Models. *ASWU = acoustically derived sub-word units.

other experiments have made use of multiple source languages (Vu et al., 2010). Since experiments have been conducted on different corpora, it is not clear which approach is superior; but both techniques work well with performance roughly at 20% word error rate on read speech. These studies and most other work on cross language transfer assumes a certain amount of knowledge about the target language, such as pronunciation dictionaries, or at the very least a phoneme inventory (Schultz and Waibel, 1998; Kohler, 1998). As far as the author is aware, only one study has looked at acoustic modelling without knowledge of the phoneme inventory; a study by Walker et al. (2003) included an experiment that made no assumptions at all about the target language. Using ten languages, a *Universal Phone Recognizer* was built with a set of 107 base phones, and trained on conversational telephone speech. This was then tested on speech in the Farsi language and resulted in a 87% PER (phone error rate). When knowledge of the inventory was included, this improved to 73% PER. Clearly, cross language phone recognition is already a challenging task, but even more so when there is minimum knowledge of the target language. This is because there are more sounds to distinguish when the phoneme inventory is not known. Ideally every possible phonetic contrast that might occur in a language needs to be detected.

Acoustically derived sub-word unit (ASWU) based modelling

There have been other attempts at language independent models that are not explicitly phone based. One approach that makes the least assumptions about the nature of speech is the use of acoustically derived sub-word units (ASWUs). These units can either be concrete exemplars of speech (Moore et al., 1983; De Wachter et al., 2007) or generalizing models (Paliwal, 1990). The units are derived directly from the speech signal; either in a supervised fashion with words labelled, or unsupervised where there is no transcription at all. ASWUs have worked well for specific tasks but as a consequence of this data-driven approach ASWUs originally tended to be speaker dependent (Ostendorf, 1999); performance dropped when moving to other speakers.

One attempt at creating a supervised speaker-independent system integrated the process of the dictionary building stage as part of the data driven process (Bacchiani and Ostendorf, 1999). This has lead to performance that is similar to phone based systems on simple tasks such as the DARPA resource management task. One attempt at a larger vocabulary problem involved a hybrid system combining both supervised speaker independent ASWUs and phone-based models (Bacchiani and Ostendorf, 1998). This was only marginally better than the baseline system and fell far short of the performance of contemporary systems that were purely phone-based (Zavaliagkos et al., 1998). Attempts on using fully unsupervised ASWUs with speaker adaptation has shown some promise but only on a small dataset (Varadarajan and Khudanpur, 2008). The difficulty of handling multiple speakers is reflected in a similar field of acoustic pattern discovery (Park and Glass, 2008); where acoustically derived units that can be longer than words are derived from repeated patterns. Most work in this area of unsupervised pattern discovery has been speaker dependent (Kempton, 2007; Park and Glass, 2008). Preliminary experiments indicates that handling multiple speakers may require a more supervised approach where utterances are given a semantic label (Aimetti et al., 2009, 2010). Moore (2007) suggests speaker normalisation could be achieved using recognition-by-synthesis where a vocal tract model attempts to mimic the input received. Units could then be derived from these motor sequences.

When performing the analysis to create ASWUs, it can be informative to inspect exactly what units have been chosen. When the algorithm is allowed to choose a small number of units e.g. equivalent to the number of phonemes in the language, often there is a rough correlation with broad phone classes e.g. approximants, fricatives, nasals (Lin, 2005, p.25,65-67), (El Hannani, 2007, p.48). When the algorithm is allowed to choose many more units, there is some evidence of a correlation with individual phone-model states and certain allophonic details (Bacchiani and Ostendorf, 1999; Varadarajan et al., 2008). With highly supervised conditions using minimal pairs on a single speaker it may be possible to identify phoneme-like units (Moore et al., 1983); but there does not appear to be any evidence of deriving the phonemes of a language with unsupervised ASWUs.

When faced with an underdocumented language, clearly there is the appeal of unsupervised ASWUs for building a recognition system, because there is not a need for transcription. However, if the units selected risk being speaker dependent and of minimal linguistic relevance, the approach is unlikely to help identify the contrastive sounds of a language.

Despite the difficulties in using ASWUs for multiple speakers; there has been a suggestion that the same ASWUs could be used across multiple languages (Chollet et al., 1999). The only area that this appears to have been attempted in is language identification of speech (see Section 2.4.2). Petrovska-Delacrétaz et al. (2003) used a common set of 64 ASWUs that were shared between Swiss French and Swiss German. However the results were not very successful partly because over 30% of the test data did not register as even containing any of the relevant ASWUs. In a more comprehensive experiment Li et al. (2007) used a set of ASWUs shared between the

languages of the 2003 NIST language recognition evaluation. This resulted in an equal error rate of 4%, of which was not quite as good as the 1.8% equal error rate for a phone-based system (Matějka et al., 2006).

These results seem to confirm that ASWUs can give good accuracy for the domain they were trained in, but outside that specific domain they appear to lack robustness and perform worse than phone-based models.

Feature-based modelling

Another approach to language independent acoustic modelling is to use linguistically-based features. The features most commonly used for this purpose are articulatory features (AFs). These describe the approximate state of the speaker's articulators and can be viewed as the components of a phone. For example the English phones [s] and [z] might be characterised as

[s]	[z]
manner = fricative	manner = fricative
place = alveolar	place = alveolar
voice = no	voice = yes

The advantage of AFs over phones is that there is a smaller number needed to characterise the different sounds in all the world's languages. They also have the potential to better model asynchronous feature spreading since a purely phoneme-based approach is slightly naive in assuming speech is just a sequence of symbols i.e. the "absolute slicing hypothesis" (Goldsmith, 1976) or the "beads-on-a-string model" (Ostendorf, 1999). AFs have been used alongside traditional phone based models to improve pronunciation variation and improve performance in noise for English ASR (Deng and Sun, 1994; Kirchhoff, 1998). However Schultz and Kirchhoff (2006, p.99) point out that "no results have yet been published showing that a recognition system solely based on alternative sound units outperforms the standard phoneme-based approach".

These articulatory features have also been used in cross-language experiments. Wester et al. (2001) trained AFs on English telephone speech, and then tested them on Dutch telephone speech. Many of the AFs transferred well between languages, but it was found that the AF for place suffered a reduction in accuracy. Stüker et al. (2003) used GlobalPhone, a broadcast quality corpus, to extend this idea further for five different languages. First AFs were trained on one language. When tested on the same language this typically resulted an average AF accuracy of 94% and when tested on the remaining four unseen languages this dropped to 87%. When the AF detectors were trained on multiple (i.e. four) languages and tested against the remaining unseen language the average AF accuracy was only slightly better at 88% (Schultz and Kirchhoff, 2006).

These cross-language feature recogniser results are difficult to compare with cross-language phone recogniser results because there is a lack of experiments where word and phone error

rates can be given. However a study by Siniscalchi et al. (2008) used binary feature detectors (AFs with only two values) as the only front end to a language specific phone recogniser. The detectors themselves give probability estimates of a binary feature being positive, and experiments showed that these binary features transferred well across languages. Training the feature detectors on five languages; and testing on a single unseen language resulted in a 48% phone error rate (PER). However a subsequent paper indicates the average error rate tested across all the target languages gives a more modest, but still competitive, 63% PER (Lyu et al., 2008).

It is difficult to make exact comparisons between all the results given for the different modelling techniques. The experiments are on different corpora, also some results are reported as phone error rates with other results reported as word error rates. Figure 2.3 is an attempt to summarise the findings comparing feature-based, phone-based and ASWU-based models. Note that if models are language dependent and / or context dependent then they are generally further down the list in the figure than the independent version of the model. For example a context dependent phone model trained on English will be more accurate on English than a context independent phone model trained on English. The latter model is likely to be more portable with the potential to be used for cross-language recognition.

2.4.2 Language identification

Language identification of speech (LID) is a field that makes minimal assumptions about the target language, often just relying on the acoustics to build the model. External knowledge sources can be used in characterising the language but these are not usually specific to the target language. Often the phoneme inventory is not assumed to be known. This blind characterisation of a language at the level of sound patterns, is highly relevant to a phonological analysis of an unknown language.

Performance of LID systems is regularly assessed by NIST (US National Institute of Standards and Technology) through the Language Recognition Evaluations which currently run about once every two years. This means that the history of improvements in language ID can be reliably recorded. The most successful language identification systems roughly split into two types *acoustic* and *phonotactic*. These are described in detail below.

Acoustic language identification

The units used in the acoustic approach are spectral frames, e.g. standard 10ms MFCC vectors. Deltas which are the difference between MFCCs for successive frames are also included. The only labelling is the name of the language; no further other training material such as transcripts are needed.

In the early-90s Riek et al. (1991), Nakagawa et al. (1992) and Zissman (1993) investigated an HMM (hidden Markov model) based approach for language identification. However it was

found that a GMM (Gaussian mixture model), which can be viewed as a single state HMM, was just as successful.

At Lincoln Labs, Zissman (1996) improved the GMM using channel normalisation but showed that it still lacked accuracy when compared with the phonotactic approach. One disadvantage with the standard GMM is that it models almost no temporal information apart from the delta cepstra. Torres-Carrasquillo et al. (2002) experimented with the previously developed SDC (shifted delta cepstra). This works by stacking delta-cepstra across multiple speech frames to model longer temporal features. This technique got performance much closer to phonotactic results. Looking for further ways to take advantage of temporal modelling, the team also produced a GMM tokeniser, where a symbol was produced for each frame corresponding to the highest scoring Gaussian component. This can be viewed as very short ASWUs. These streams of tokens were then evaluated by n-gram language models. However the gain was marginal and only gave a small improvement when fused with the existing GMM acoustic scores.

Developing the Lincoln Labs entry to the NIST 2003 LID evaluation, Singer et al. (2003) used the SDC GMM, and with improved channel and gender normalisation and achieved a result of 4.8% equal error rate (EER) in the main competition. This was the first time a GMM system had surpassed the performance of a phonotactic model. A different acoustic method was also attempted on the data. Following Lincoln's successful use of the SVM (support vector machine) in speaker recognition, the same approach was used on the NIST language recognition data. The same feature vectors were used resulting in 6.1% EER, (although on an older dataset the SVM had been slightly superior). Campbell et al. (2006) reports that a fused system of the GMM and SVM on the primary NIST test condition resulted in 3.2% EER indicating that the two modelling systems provide complementary information.

The GMM is traditionally trained using Maximum Likelihood training. In LVCSR (large vocabulary continuous speech recognition systems) there has been a lot of success in replacing the maximum likelihood training with discriminative training. At Brno, Burget et al. (2006) experimented with discriminative training on the GMM trying MCE (minimum classification error) and MMI (maximum mutual information). The latter was the most successful, reducing the number of mixture components from the 2048 in the traditional GMM to just 128 mixture components in the MMI-trained GMM. This simpler model gave an improved result of 2.0% EER. A few other smaller improvements were made, including the use of HMMs and data decorrelation but these were less significant.

Acoustic-based language identification has historically borrowed many techniques from the larger field of speaker identification. This is particularly the case with Joint Factor Analysis (Kenny et al., 2007). The technique attempts to separate the channel/session variation (unwanted) with the speaker variation (wanted). When applied to language identification this showed competitive results on the NIST 2009 evaluation (Jančík et al., 2010). This alongside other LID results are shown in Table 2.2.

System	Reference	Test-30s	Error
Acoustic GMM	Singer et al. (2003)	NIST-2003	4.8% EER
Acoustic SVM	Singer et al. (2003)	NIST-2003	6.1% EER
Phonotactic P-PR-LM	Singer et al. (2003)	NIST-2003	6.6% EER
Phonotactic PR-LM	Matějka et al. (2006)	NIST-2003	1.8% EER
Acoustic GMM-MMI	Burget et al. (2006)	NIST-2003	2.0% EER
Acoustic ASWU-VSM	Li et al. (2007)	NIST-2003	4.0% EER
Phonotactic UPR-VSM	Tong et al. (2009)	NIST-2003	1.42% EER
Feature/Phonotactic VSM	Siniscalchi et al. (2010)	NIST-2003	8.5% EER
Phonotactic PR-SVM	Jančík et al. (2010)	NIST-2009	1.78 C_{avg}
Acoustic GMM-JFA	Jančík et al. (2010)	NIST-2009	2.02 C_{avg}

Table 2.2: Phone-based language identification has generally shown superior performance

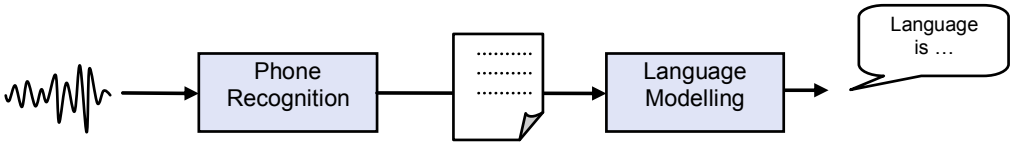

Figure 2.4: An example of phone-based language identification: Phone Recognition followed by Language modelling (PR-LM)

Phonotactic language identification

One type of system that has consistently performed well through all the NIST evaluations is the phonotactic approach to LID. This technique splits the problem into two stages; phone recognition and then language modelling (PR-LM) of the tokenized phones (see Figure 2.4). In one of the earliest studies on LID, House and Neuburg (1977) laid the groundwork for this approach. They investigated language modelling on phone transcripts because at that time phone recognisers for audio were not accurate enough. Phones were clustered into four broad classes and HMMs were used for the language modelling.

Hazen and Zue (1993) took this work further using real phone recognisers on audio data. These phone recognisers had to be trained from annotated transcripts. Phones were clustered into broad classes and n-grams were used for the language modelling. Hazen and Zue experimented with different phone classes, looking both at unsupervised and manual selection of phone classes. It was found that 23 manually selected phone classes, based on the clustering of English phones worked best. Overall they found that the PR-LM model performed better than acoustic or prosodic models.

In similar work, Zissman and Singer (1994) also found that the acoustic approach was surpassed by PR-LM performance and they continued to experiment with different variants of PR-

LM. They started with 48 English phones for the recogniser which they found worked better than coarser groupings. This meant that other languages were being transcribed into English phones, but separate language models could still be built on the output of the English phone recogniser. In fact this was the main advantage of keeping the PR and LM stages separate; it was possible to build language models for languages where no phone annotations existed for training phone recognisers. If phone annotations are available then there isn't the same requirement to keep the PR and LM stages separate and there can be advantages in having a more tightly coupled system i.e. the phone recogniser can benefit from the phonotactic constraints of the language model to make decisions. Each language can then have its own integrated model. Originally proposed by Lamel et al. (1993) and Muthusamy (1993), Zissman and Singer called this PPR (parallel phone recognition) and showed that this approach worked well.

When phone annotations are available in more than one language, they can be used to build multiple PR-LM systems, e.g. as well as an English phone recogniser producing a number of language models, a German recogniser might be doing the same. Zissman and Singer call this system Parallel-PR-LM and the performance is similar to PPR with the added flexibility of using it for languages where annotations are not available.

There have been studies looking at whether the phone recognisers for new languages can be built without phone transcripts. Tucker et al. (1994); Lamel et al. (1994); Lloyd-Thomas et al. (1998) have investigated adapting a phone recogniser from one language to another using a bootstrapping approach but this has had limited success.

Many phonotactic approaches use language dependent phones for Language ID, however there are advantages in using language independent phones such as the ability to use discriminant training between multiple phone transcripts. These multi-lingual phone sets include broad classes as used by Muthusamy (1993), or more finely grained phones picked to maximise the discrimination between particular languages (Berkling, 1996). Li et al. (2007) used multilingual ASWUs as tokens in a mix of acoustic and phonotactic methods. This worked well but the language dependent phone-based approach still performs the best.

With the addition of duration modelling, better silence and closure modelling, Singer et al. (2003) evaluated the Lincoln six-level Parallel-PR-LM system on the NIST 2003 dataset resulting in a 6.6% EER.

At Brno, Matějka et al. (2006) took the PR-LM model and worked on building a more accurate phone recogniser, this was done through large volumes of additional labelled training data and a different design based on neural networks rather than HMMs. A single-level PR-LM with a Hungarian phone recogniser (62 phones) achieved 3.1% EER. Further improvements were made by using a phone lattice developed earlier by Gauvain et al. (2004). Matějka also added anti-models which are trained on misrecognised segments. This improved performance to 1.8% EER.

Tong et al. (2009) produced a slightly better result using a universal phone recogniser. This

was based on 300 phones lumped together from different languages which was then pruned to a subset that was particularly suitable for making discriminations for the target language.

In recent years language identification has been dominated by the phonotactic approach. For example at Odyssey 2010 (the speaker and language recognition workshop) and Interspeech 2010, the clear majority of papers described systems that were phone-based. Jančík et al. (2010) showed that state-of-the-art phonotactic approaches outperform state-of-the-art acoustic approaches.

There has been research into other units to use in LID. Syllables have been studied a number of times, but work by Martin et al. (2006) shows that they are not as competitive as a phone based approach. Prosody has been found to be good at distinguishing pairs of languages but not found to work well over a range of languages (Zissman and Berkling, 2001). On articulatory features (AFs) Parandekar and Kirchhoff (2003) created a system that used a feature recogniser followed by a feature-based language model. This worked better than a phone-based baseline but, since the baseline did not perform as well as state-of-the-art phone-based LID systems and the test conditions were not quite the same as NIST, the results are inconclusive. Siniscalchi et al. (2010) used two AFs; manner and place which with a discriminative classifier gave a result of 8.5% EER. When compared to systems trained on the same amount of small data, the result is more competitive than may first appear.

LVCSR-based language ID performs well but this requires comprehensive characterisation of the language in the first place (Schultz et al., 1996). Since the challenge of low resource languages is often reflected in the NIST evaluations, this technique is not a serious contender.

A summary of language identification results are shown in Table 2.2.

The surprising element in language identification is how well the phonotactic approach works. Given that a fully acoustic approach has the potential to globally optimise the problem, it might be expected to be increasingly leading the way. However, splitting the stream of sound into phone-like units appears to be particularly good at characterising and discriminating languages. Kempton and Moore (2008) showed that by simulating very accurate phone recognizers, the phonotactic approach worked very well even when using a small number of phone categories.

2.4.3 Which speech technology to use?

To help with a phonemic analysis a phone recognizer is needed that, like an expert phonetician, can ideally detect all possible contrasts. Most speech recognition research assumes that a phoneme inventory is already known only the study by Walker et al. (2003) attempts cross-language phone recognition without this knowledge. The high error rates indicate how challenging this task is.

Figure 2.3 summarises performance of the three main types of multilingual acoustic mod-

elling. ASWUs, when compared to phone-based modelling, do not perform as well when there is a difference between test and training conditions. This indicates that ASWUs are not yet mature enough to be used in a cross-language context. Feature-based acoustic modelling has not outperformed phone-based modelling but there are strong indications that they can work well in cross-language situations.

Language identification does not have exactly the same aims to characterise a language as a phonemic analysis, but it does share the lack of assumptions the target language's phonology. In the field of language identification, there are similar findings to the rest of speech recognition. Phone-based modelling outperforms acoustic-based modelling, and feature-based modelling although not outperforming phone-based modelling shows some promise of improvement.

The robust performance of phone-based models and the wealth of resources available such as multiple language recognisers (Schwarz et al., 2009), makes phone-based models a preferred choice for use in machine-assisted phonemic analysis.

2.5 Selected areas in speech development

2.5.1 Perception of speech categories

In Chapter 1, it was shown that contrasts in one language, such as Sesotho, are not recognised in other languages such as English and vice-versa. There can also be a difference between languages where the point of contrast is in a slightly different position on a phonetic continuum. A simple example is in the perception of stops such as /b/ and /p/. The chief difference between these sounds is a difference in voice onset time; the time taken for the vocal folds to start vibrating after the stop has been released. For English word-initial bilabial stops if the voice-onset-time is less than 25ms then most English listeners perceive the test word as /ba/, if the voicing comes later it is perceived as /pa/ (Lisker and Abramson, 1970). These phenomena come under the more general psychological area of *categorical perception* (see Harnad (2003) for a contemporary definition). The precise voice-onset-time for discriminating sound categories such as bilabial stops can vary across languages and some comparisons are shown in Table 2.3. Note that all occurrences of Spanish refer to Latin American Spanish. Thai has two boundaries because there is a distinction between voiced, voiceless, and aspirated stops (/b/,/p/,/p^h/).

Williams (1977) improved on earlier Latin American Spanish measurements (Lisker and Abramson, 1970) by ensuring speakers were monolingual. Interestingly, as well as the -10ms boundary among the Spanish listeners, Williams also found a lower but significant discriminatory peak located at the same position as the English peak. Pearce (2007) compared the languages Thai, English, Kera and French showing that the VOT points are not fixed but depend on pitch. If the pitch is higher the perception and production timings tend towards a longer VOT.

Subjects	Reference	VOT (-)	VOT (+)
English adults	Lisker and Abramson (1970)		25ms
Spanish adults	Williams (1977)	-10ms	
Thai adults	Lisker and Abramson (1970)	-30ms	35ms $\pm$5ms
Infants (English)	Eimas et al. (1971)		30ms $\pm$10ms
Infants (Spanish)	Lasky et al. (1975)	-45ms $\pm$15ms	45ms $\pm$15ms
Chinchillas	Kuhl and Miller (1978)		25ms

Table 2.3: Voice onset time perceptual boundaries of word initial bilabial stops; evidence of innate boundaries at approximately -30ms and +30ms

These measurements show just how much difference there can be in perception between adults in different language groups. It is also interesting to compare these findings with infants.

One of the surprising findings in infant speech perception is that infants can discriminate subtle sound categories from a very early age. Results from two studies on infant speech perception are also shown in Table 2.3. The main difference between experiments is that Lasky et al. (1975) included pre-voiced samples; so actually the results are very similar. Summarising the findings in categorical perception for infants, Eimas et al. (1987) states that "The overall data indicates that infants divide the voice-onset-time continuum into three categories, with boundaries situated approximately at values that correspond to the voiced-voiceless boundary in English and other languages and to the prevoiced-voiced boundary of Thai, among other languages"

Categorical perception of these speech sounds is not only confined to our species. Kuhl and Miller (1978); Kuhl (1981) found chinchillas had discrimination peaks in the same location as English adult subjects for post-voicing in bilabial, alveolar and velar stops. It is difficult to judge the effect of chinchillas exposure to any previous English. This does not appear to be raised as a problem in the literature and it was also stated in the original paper that the chinchillas had originally been kept at the Institute of the Deaf which may have minimised this exposure.

2.5.2 Learning sound categories

Strict categorical perception is not observed for all speech sounds. For example there is a weaker discrimination within the vowel space which Kuhl (1991) refers to as the *perceptual magnet effect*. By six months, infants showed a better discrimination at the edge of the vowel categories than within the categories for the vowels in their language. However, the same experiment with monkeys did not show such an effect. This indicated that the phenomena was unique in humans and was learnt from exposure to spoken language Kuhl (2004). Computational simulations such as Bayesian modelling have reproduced the effect suggesting this behaviour is a consequence of optimally solving the problem of category membership and subphonemic cues (Feldman and

Griffiths, 2007).

Kuhl hypothesises that the infants are learning from the distribution of speech sounds in the ambient language i.e. exposure to a frequently occurring vowel in the native language leads to perceptual magnet effect for that vowel. Kuhl cites another study by Maye et al. (2002) for evidence of learning statistical distributions. In this study six and eight month old infants were exposed to eight different sounds varying from a [d] with prevoicing, to a voiceless unaspirated [t]. English adults tend to interpret all of these sounds as /d/ but infants maintain some discrimination until 10-12 months (Pegg and Werker, 1997). One group of infants were exposed to a bimodal distribution where there were more examples of sounds at the two extremes. The other group of infants were exposed to a unimodal distribution with more examples in the mid-range. The group trained on the bimodal sounds showed an ability that was superior to the unimodal group in discriminating bimodal test sounds.

Infants also show a sensitivity to the sequential patterns of sound. Jusczyk et al. (1994) exposed infants to monosyllable nonsense words. They showed that 9-month old (but not 6-month old) infants showed a preference for listening to phonetic patterns that matched their native language. Saffran (2003) has shown in a number of experiments that infants are able to segment word-like units just by tracking the transition probabilities between syllables. Prosody is also a helpful cue for word segmentation (Johnson and Jusczyk, 2001), playing a more important role in 8-month old infants.

Studies comparing sound discrimination and word learning introduce some puzzles. Stager and Werker (1997) found that 14-year old infants could quickly learn words that were phonetically different but they had difficultly learning and distinguishing words that were minimal pairs, even though they could already discriminate the particular sound in a speech perception experiment, and could use well-known words with the same phonemic difference (Fennell and Werker, 2003). Werker believes that the word learning task is so computationally intensive that there are not the attention resources for learning words that are so phonetically similar.

In conclusion; during the first year of their life, infants show an ability to perceive different sound categories from many different languages but as they grow up in their language environment the speech perception becomes more language specific. Babies move from being *citizens of the world* to *culture bound* listeners (Gopnik et al., 1999; Kuhl, 2004). Phonetic boundaries shift and become more defined so that, for the developed speaker, they can be in slightly different positions across different languages.

It is difficult to know how much awareness there is of phonological elements in infants. Some have argued that infants perceive larger units and it's only when they are older that there is some type of perceptual reorganization in the infant's lexicon (Lin, 2005). It could be speculated that this perceptual reorganisation might cover similar levels as in Figure 2.3. More research is needed to understand how humans developed a point where they can identify the contrastive sounds in a language. However, even a small understanding of the learning process

can identify some of the principles involved such as learning from sound category frequentness and sequential patterning.

2.6 Chapter summary

A statistical analysis of data from a previous study but completed as part of the work for this thesis confirms that literacy is related to language vitality. Survey figures indicate that many (maybe half of all) languages remain to be written. There is therefore a clear need for phonemic analysis. With the death of taxonomic phonemics, there was less optimism in automating the process and yet most of the process is still fairly mechanical and there is a demand among linguists for tools that can speed it up. There are some good tools available to help, but these just provide database functionality, and there is an opportunity for machine assisted analysis. Some progress has been made in computational phonology but not in a specific area of linking acoustics with the unsupervised discovery of allophonic rules. The least robust technology for cross-language speech recognition is acoustically derived sub-word unit (ASWU) based modelling. Feature-based modelling shows some promise but phone-based modelling is the most robust. Splitting the problem into phone recognition followed by analysis of the sequence of phones strikes a balance between tractable modelling and adequate characterisation of the sound patterns of a language. Studies in human language perception make it clear that phonetic boundaries can be in slightly different positions across different languages. However there is also some evidence that infants start off sharing the same perception of phonetic categories, and these adapt to the language that they are learning. The learning process includes the influence of sound category frequency and sequential patterning.

Chapter 3

Phonetic similarity

It was stated in Section 1.2 that, in a phonemic analysis, relying on some notion of phonetic distance is sometimes implicit but always important. In practice when performing an analysis many linguists will make the assumption that some sounds are too phonetically dissimilar to be allophones e.g. [m] and [k] (Gleason, 1961, p.275). However, many authors are deliberately cautious in defining any universal threshold of phonetic similarity (Hayes, 2009, p.54; Clark et al., 2007, p.97). Pike, instead of defining phonetic similarity by a rule, illustrates the principle through examples of possible allophone pairs covering over 100 different sounds based on his experience of phonemic analysis (Pike, 1947, p.70).

In this chapter different phonetic distance heuristics are evaluated *quantitatively* for their effectiveness in detecting allophones. Up to now this has received little attention in the literature. The structure of the chapter is as follows. Section 3.1 covers the evaluation of phonetic similarity detection algorithms (where an algorithm classifies a pair of phones as either similar or dissimilar). Section 3.2 covers the more generalised case (where an algorithm gives a distance), and also introduces one of the principle evaluation metrics used in this thesis. Section 3.3 deals with the comparison of sound sequences. In these three subsections the experiments were performed on the English language in such a way as to simulate an under-resourced language. Section 3.4 introduces a number of additional corpora suitable for testing machine-assisted phonemic analysis. One of these is the Kua-nsi language which is then used for an additional evaluation of the algorithms introduced in this chapter. Section 3.5 gives the conclusions.

3.1 Phonetic similarity detection

3.1.1 Relative minimal difference (Experiment 3A)

Peperkamp et al. (2006) makes use of phonetic similarity in an algorithm to model the acqui-

sition of allophonic rules by infants. The main algorithm attempts to detect allophones via complementary distribution by measuring discrepancies in context probabilities for each pair of phones. This is investigated further in Section 5.3. Peperkamp also introduces *phonetic filters* acting as a post process after the main algorithm to remove spurious allophones i.e. pairs of phones that are not actually allophones but are phonemically distinct. One of these filters makes use of phonetic similarity to reject spurious allophones. A *minimal distance* criterion is formalised, where a pair of phones are judged to be spurious allophones if there are any other phones between them in phonetic space; "for each of the [phonetic features], the third [phone] lies within the closed interval defined by the other two" (Peperkamp et al., 2006). In this thesis Peperkamp's minimal distance is referred to as the *relative minimal difference* to avoid confusion with similar terms; the word *relative* is used to indicate that any prediction of an allophonic relationship is affected by the presence of other phones in the phone set. For example, if the only glottal fricatives to appear in a transcription are [h] are [ɦ] then these are judged as possible allophones because there are no other sounds in the transcription phonetically between them.

It was decided that this implementation of phonetic similarity could be more fully evaluated and compared with other measures, which is the subject of this chapter. In the original study (Peperkamp et al., 2006), this relative minimal difference algorithm helped to detect allophones when combined with other algorithms, but it was not tested by itself. In this chapter Peperkamp's phonetic similarity is evaluated for its effectiveness as a standalone process.

Phonetic representation

Peperkamp et al. (2006) used five multi-valued articulatory features to represent French speech sounds. However, the particular articulatory features framework is not expressive enough for many other languages. For the current work it was decided that an all-binary feature system would be more suitable. The main appeal of binary features is their simplicity for algorithmic implementation and their flexibility in representing speech sounds with multiple articulations. For example, a labial-velar approximant [w], a velarized lateral [lˠ] and an r-coloured vowel [ɚ] cannot be fully defined with the multi-valued features used in Peperkamp et al. (2006), but they can with binary features. There are also many practical resources available for using binary features e.g. Hayes (2009) specifies a universal set giving definitions for 141 phones that can be easily extended to other sounds; 28 binary features are defined and most of these features are included in Figure 3.1. These resources are available online and are used in the experiments for this thesis. It is also possible to add further features such as tone. There are many practical advantages to using binary features, but some of their theoretical shortcomings are described in Section 3.2.3.

Corpora for evaluation

In both Peperkamp et al. (2006) and a follow-up experiment (Le Calvez et al., 2007) the algorithms were tested on a corpus of child directed speech. Originally this corpus was transcribed as text, but for their experiments it was automatically converted to a phonemic transcription and allophones were added with predefined rules.

In the initial experiments in this thesis, the algorithms of Peperkamp et al. were evaluated on the TIMIT corpus; a dataset that contains allophones that have been labelled manually directly from the acoustic signal. This means the transcript used here is more faithful to the acoustics than in the previous published experiments. The TIMIT corpus (Garofolo et al., 1993) of US English was chosen because it is one of the largest corpora available that contain manually annotated allophones. The TIMIT transcripts of 1386 utterances were used as evaluation data in subsequent chapters of this thesis. For this chapter on phonetic similarity, it is only the phone set (derived from the phonetic transcript) that is needed.

A number of different sources (Garofolo et al., 1993; Esling et al., 1999; Hieronymus, 1993) were used to confirm the conversion of the TIMIT symbols to IPA. The sound /r/ is known to have a number of realisations in US English e.g. [ɹ, ɹ̝, ɻ]; in this experiment the retroflex approximant [ɻ] is used because it shares a number of binary features with the other realisations. The analysis was restricted to consonants for the purpose of starting with a simple problem that did not involve phone sequences such as diphthongs. The problem of phone sequences is investigated later in Section 3.3. All the US English consonant sounds had feature definitions in Hayes (Hayes, 2009) except for the nasalized alveolar tap [ɾ̃]. The features defined for this phone were the same as the alveolar tap [ɾ] but included the feature [+nasal]. The full list of consonants is shown in Figure 3.1.

Evaluation measure and results

The outcome of Peperkamp's relative minimal difference criterion applied to the TIMIT consonants is shown in Figure 3.2. This *phone relationship chart*, has the same layout as triangular mileage charts on street atlases which have the cities listed on a diagonal. Here phones are listed in place of cities, and relationship between the phones listed in place of the mileage. The phone relationship chart is introduced in this thesis to give a visual representation of phone relationships e.g. showing whether two phones are phonemically distinct or are allophones of the same phoneme, and to show other relationships that might predict this. To make the best use of space, two different triangular charts have been combined into a square. In this figure the bottom left triangle corresponds to the relative minimal difference criterion, and the shaded cells containing a one indicate that the phone pair may be in an allophonic relationship. The top right triangle corresponds to a different algorithm described in Section 3.1.2. Cells with an outline show the ground truth where a phone pair has an allophonic relationship according to

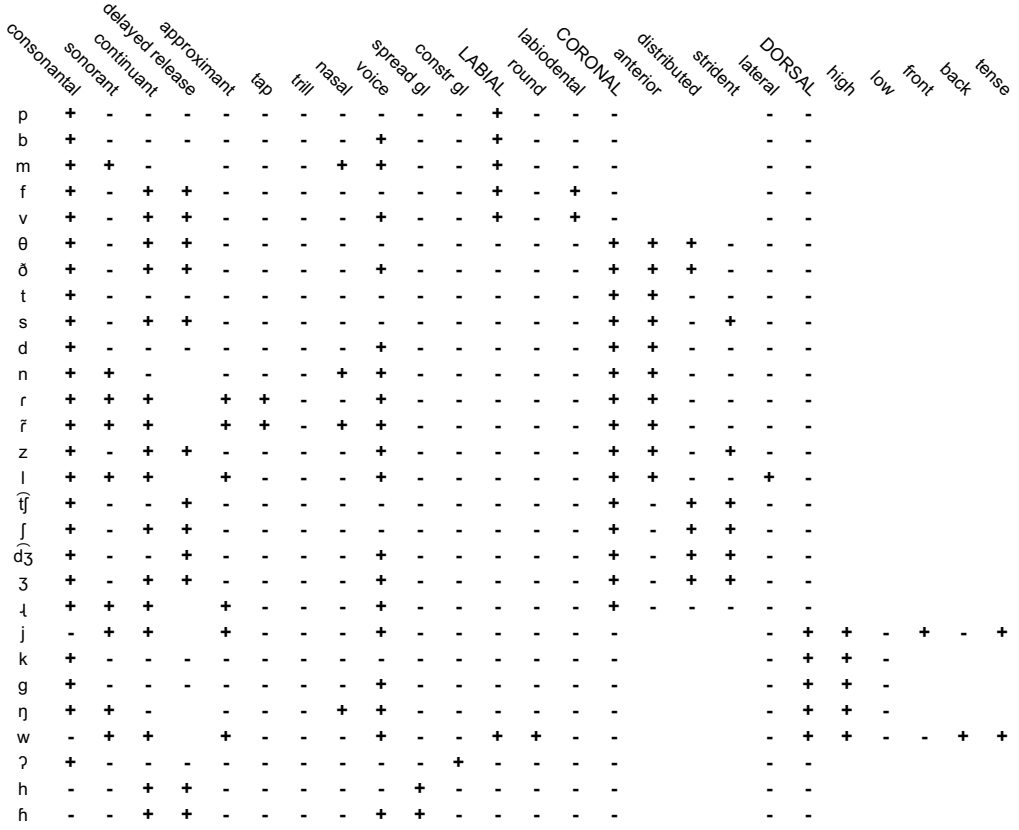

Figure 3.1: TIMIT consonant features (includes redundancy)

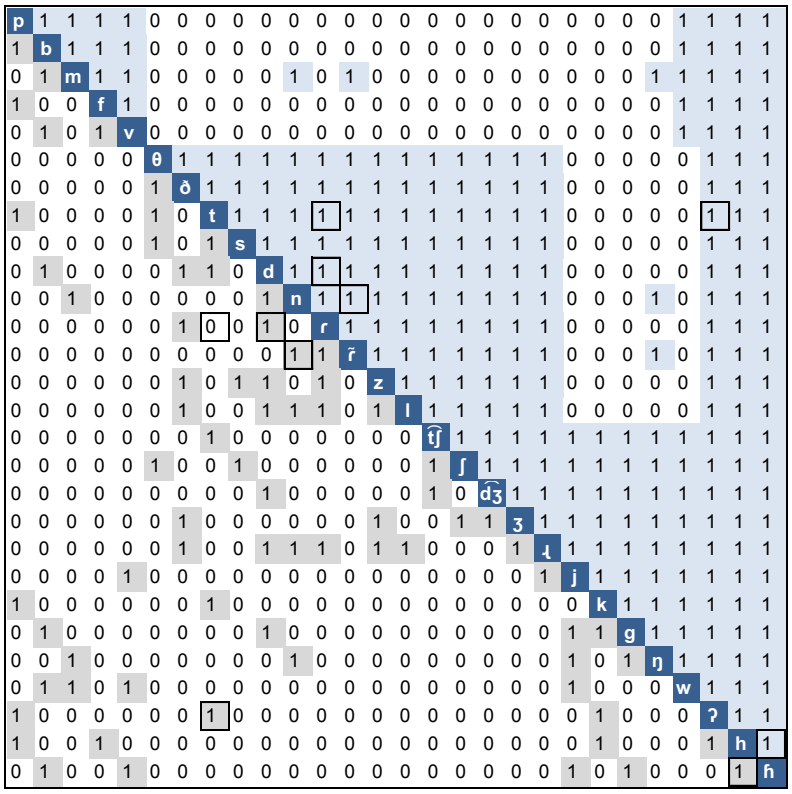

Figure 3.2: Phone relationship chart showing the relative minimal difference based detector (bottom, Experiment 3A) and active articulator based detector (top, Experiment 3B). Each cell represents a phone pair that, if marked as '1' and shaded, is judged as a possible allophone. Outlines mark actual allophones.

the TIMIT documentation. Figure 3.3 shows the example where [t] and [n] are not judged to be allophones. This is because [d] lies phonetically between the two; i.e. it is both [+voice] and [-nasal]. The relative minimal difference criterion correctly rejects this phone pair.

Equivalent results of Figure 3.2 are shown in Table 3.1. The standard information retrieval measures given are derived from the following contingency table:

	Present	Not present
Detected	hit	false alarm
Not detected	miss	correctly rejected

$$\text{False alarm rate} = \frac{\text{false alarms}}{\text{false alarms} + \text{correctly rejected}}$$

$$\text{Recall} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

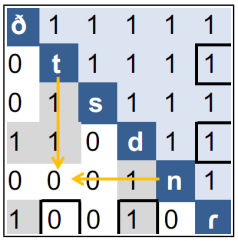

Figure 3.3: Phone relationship chart (detail) showing that phones [t] and [n] are not predicted to be allophones (Experiment 3A).

	Rel. min. difference	Articulator	Combined
Hits	4	5	4
Misses	1	0	1
False alarms	69	233	62
Correctly rejected	304	140	311
False alarm rate	0.185	0.625	0.166
Recall	0.80	1.00	0.80
Precision	0.0548	0.0210	0.0606

Table 3.1: Results of both the allophone detectors (Experiment 3A and 3B)

$$\text{Precision} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

A recall of 80% (0.8) reflects one erroneous miss of the apparent allophone pair [ɾ,t]. This is because [d] lies phonetically between the two. Although this is a slightly complex case involving both allophony and neutralisation, this could also be an issue whenever there are multiple allophones (e.g. in Maasai, [k, g, ɣ] are all realizations of /k/ (Hayes, 2009, p.39)). It is possible that this filter could be run more than once after the phone inventory is updated, but the overall process would need to have a high accuracy of detection with a suitable stopping condition. Le Calvez et al. (2007) also recognise this problem of detecting multiple allophones and suggest a modification to the relative minimal difference algorithm. When searching for sounds that could be between a pair, the set of sounds searched for are restricted to just those sounds that appear in the context of (i.e. next to) the hypothesised allophone. This solves the problem with the above example because [d] never occurs in the context of the [ɾ] allophone. This means that [ɾ,t] are now directly detected as an allophone pair. However, this modification could cause problems with other languages that have very simple syllable structures. If a language has only CV syllables, then when comparing consonants the immediate context would be a vowel. This would lead to the majority of consonant pairs being mistakenly labelled as allophones.

3.1.2 Active articulator (Experiment 3B)

A new phonetic similarity detection algorithm is introduced that draws its inspiration from linguists. This is based on the articulators that are used. Linguists involved in phonemic analysis use a number of guidelines to narrow down the number of comparisons that need to be made between phones. In a similar way to Pike (1947, p.70), Burquest (2006, p.51) shows graphically which sounds can be considered similar and these are generally orientated around different active articulators. The heuristics used by Burquest are from a perspective of marking possible allophones. Here, some of these heuristics are reinterpreted from the opposite perspective of predicting whether or not two phones are phonemically distinct. The generalised heuristic is that if two phones use distinctly different active articulators, then it is predicted that the phones are phonemically distinct.

This can be described more formally as follows. A set of active articulators is defined which includes the lips, tongue and velum i.e. the binary features: {labial, coronal, dorsal, nasal}. A dorsal coronal overlap element is also included because there can be overlap in the postalveolar and palatal region (e.g. in some languages [t͡ʃ] is an allophone of /k/ (Burquest, 2006, p.54)):

+dorsal → dorsal_coronal_overlap
+coronal, –anterior → dorsal_coronal_overlap

So the active articulator universal set is:

$$U_{\mathrm{AA}} = \{\text{labial, coronal, dorsal\_coronal\_overlap, dorsal, nasal}\}$$

The active articulator set of each phone can include any number of these possibilities.

$$a, b \subseteq U_{\mathrm{AA}}$$

Here a,b represent the active articulator set used by the different phones. Phonemic distinctiveness is predicted if both phones are using distinctly different active articulators, i.e. the following three conditions are all met.

$$a \neq \emptyset$$

$$b \neq \emptyset$$

$$a \cap b = \emptyset$$

Example 1, comparing [p] and [t]:

$[\mathrm{p}]_{\mathrm{AA}} = \{\text{labial}\}$
$[\mathrm{t}]_{\mathrm{AA}} = \{\text{coronal}\}$
$[\mathrm{p}]_{\mathrm{AA}} \cap [\mathrm{t}]_{\mathrm{AA}} = \emptyset$

All the conditions are met, therefore [p] and [t] are predicted to be phonemically distinct.

Example 2, comparing [k] and [ʔ]:

$[k]_{AA}$ = {dorsal, dorsal_coronal_overlap}

$[ʔ]_{AA} = \varnothing$

The second condition is violated, therefore [k] and [ʔ] are predicted to not necessarily be phonemically distinct.

Example 3, comparing [n] and [ŋ]:

$[n]_{AA}$ = {coronal, nasal}

$[ŋ]_{AA}$ = {dorsal, dorsal_coronal_overlap, nasal}

$[n]_{AA} \cap [ŋ]_{AA}$ = {nasal}

The third condition is violated, therefore [n] and [ŋ] are predicted to not necessarily be phonemically distinct.

Overall this heuristic is relatively conservative in predicting phonemic distinctiveness and more liberal rules could be stated, although the rules may have to be expressed slightly differently for different feature systems. This particular phonetic similarity criterion is not a relative measure like Peperkamp's because it doesn't need to take into account other sounds observed in the language. The results of this active articulator filter applied to the TIMIT consonants is shown on the top right side of Figure 3.2 and the results in Table 3.1.

For the active articulator based detector it can be seen that there are no misses, but many false alarms leading to 100% recall and low precision. This characteristic of high recall is valuable when it is important not to miss any allophones. An investigation of the French and Japanese phonetic data in Peperkamp et al. (2006) and Le Calvez et al. (2007) reveals that this active articulator algorithm would also not miss any allophones in these languages either.

Figure 3.2 indicates that there is some correlation between the relative minimal difference and active articulator algorithm, but that they also complement each other. Table 3.1 includes a combined algorithm result which is the same combination method as Peperkamp et al. (2006). Peperkamp combines the allophone detection algorithms by viewing them as stacked filters to only allow allophone pairs through. This can be viewed as a logical AND combination, or if the values are represented as numerical scores, as a simple multiplication. The result shows that the combined algorithm has a slightly higher precision.

3.2 Phonetic distance measure (Experiment 3C)

Another heuristic that linguists use for phonetic similarity, is the number of features that differ, although it is acknowledged that not every single feature will carry the same weight of salience

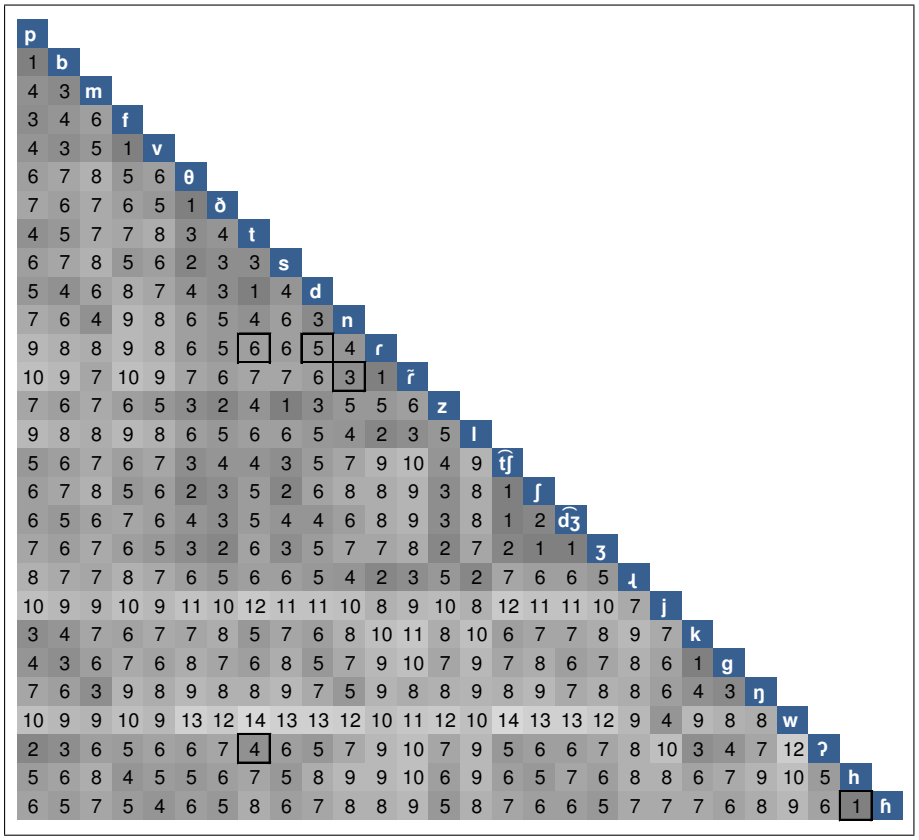

Figure 3.4: Phone relationship chart showing how many binary features the phones differ by (Experiment 3C). Darker shading indicates that the two phones are similar. Outlines mark actual allophones.

(Burquest, 2006, p.52). In the absence of a universal comprehensive account of the different features and their salience, it is helpful to start with a simple algorithm; the binary feature distance.

Given the pair of sounds [k] and [ŋ], a number of different features can be counted. It can be seen from Figure 3.1 that there are three different binary features (nasal, sonorant, voice) and one feature (delayed release) is underspecified for the sound [ŋ]. Underspecified features are assigned a value halfway between the corresponding +/- values. Therefore, in this case the distance is 3.5 binary features. Figure 3.4 shows this measure for all the non-syllabic consonants in TIMIT with distances rounded to the nearest whole number. The shading used in this chart is consistent with all the two dimensional charts in this thesis, where a darker shade represents a similarity and a lighter shade represents more of a difference.

3.2.1 Evaluation measure and results

A threshold can be chosen for the binary feature distance, e.g. it might be decided that any value less than six could be a possible allophone. But rather than choosing a particular threshold, the scores were kept in a ranked list allowing the threshold to be chosen by the linguist. The performance of the ranked list was measured using two information retrieval summary statistics. The first is ROC-AUC (receiver operating characteristic - area under curve). This can be derived by plotting a graph of recall against false alarm rate, and measuring the area under the curve. An example can be seen in Figure 3.5. The measure can also be interpreted as the probability that a randomly chosen target (allophone pair) will have a higher score than a randomly chosen non-target (non-allophone pair) (Bamber, 1975). For example a randomly ranked list will have a ROC-AUC of 50% and a perfectly ranked list will have a ROC-AUC value of 100%. The figures for the binary feature distance measure in Table 3.2 show it is having a beneficial effect.

The second information retrieval statistic is PR-AUC (precision recall - area under curve). This can be derived by plotting a graph of precision against recall and measuring the area under the curve. An example can be seen in Figure 3.6. It is a very similar measure to average precision which is widely used in information retrieval literature and at TREC information retrieval evaluations. Aslam and Yilmaz (2005) show that PR-AUC (which they call actual average precision) is strongly correlated to average precision, and suggest it may be better for evaluating the quality of the underlying retrieval function. For completeness the average precision is given alongside PR-AUC in Appendix A. PR-AUC gives a different view on performance to ROC-AUC, and it is orientated towards the perspective of the linguist; representing an expectation of precision where precision can be viewed as the probability of detected targets in the ranked list. It is affected by the proportion of targets in the original data set, which means it is not suitable for comparing results across datasets. For example a randomly ranked list of all the possible phone pairs in TIMIT would have a PR-AUC value of 1.3%, whereas a randomly ranked list of the Kua-nsi dataset introduced in Section 3.4.3 would have a PR-AUC value of 0.7%.

The ROC-AUC measure of performance is from the perspective of targets present in the original data set. It is not affected by the original proportion of targets and is suitable for comparing results across datasets. That is why a randomly ranked list has a ROC-AUC value of 50%, whatever the dataset. The ROC-AUC statistic should be therefore regarded as the primary evaluation measure.

ROC-AUC and PR-AUC were calculated[1] with AUCCalculator (Davis and Goadrich, 2006). Average precision was calculated with the Trec_eval tool (Buckley et al., 2006). One issue with

[1]When a ranked list is reversed the ROC-AUC should be 100% minus the ROC-AUC of the original list. There is currently a bug in the software AUCCalculator 0.2 available at http://mark.goadrich.com/programs/AUC/ which can give incorrect results in this case. This bug was discovered during preliminary experiments for this thesis. A corrected version generously provided by the authors was used in the experiments here but at the time of writing this has not been released on their webpage.

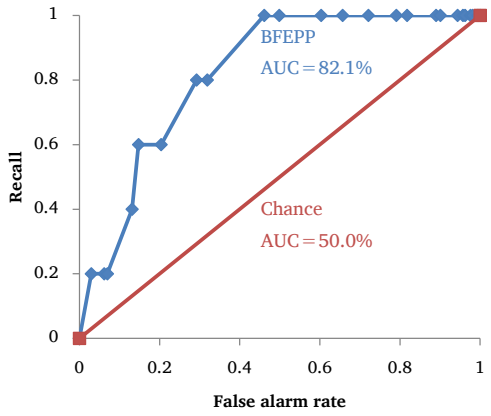

Figure 3.5: Receiver operating characteristic graph showing area under the curve (ROC-AUC) for binary feature distance and chance.

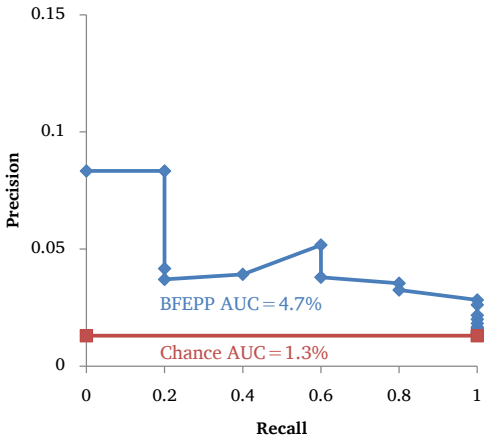

Figure 3.6: Precision-recall graph showing area under the curve (PR-AUC) for binary feature distance and chance.

Algorithm applied to TIMIT	ROC-AUC	PR-AUC
Binary feature distance	82.1%	4.7%
Relative minimal difference	80.8%	5.1%
Active articulator	68.8%	2.1%

Table 3.2: Area under the ROC and PR curves for the different algorithms on TIMIT, for Experiments 3C, 3A, 3B

Trec_eval is that any items with the same score are given a deterministic but arbitrary ranking (based on the label text) before calculating average precision. This means that results can change slightly depending on the Unicode allophone label. PR-AUC is not affected by this issue which is another reason why it is preferred to average precision. ROC-AUC is also unaffected and is the primary evaluation measure.

3.2.2 Comparison with phonetic similarity detection algorithms

The probabilistic interpretation of ROC-AUC given earlier allows an equivalent figure to be calculated for the phonetic similarity detection where there are only two scores (0 and 1). ROC-AUC as a single performance measure allows different algorithms to be compared. Results for the distance and detection algorithms investigated are shown in Table 3.2. The method for calculating ROC-AUC when there is a detection threshold, i.e. only two scores, is as follows:

$$\text{ROC-AUC}_{point-value} = r(1-a) + \tfrac{1}{2}ra + \tfrac{1}{2}(1-r)(1-a)$$

where:
r = recall (true positive rate)
a = false alarm rate (false positive rate)

The rationale behind this is that randomly choosing a target (allophone pair) and the non-target (non-allophone pair) are two independent events. With a single threshold there are four possible outcomes.

1. Target is above threshold, non-target is below threshold
2. Target is above threshold, non-target is above threshold as well
3. Target is below threshold, non-target is below threshold as well
4. Target is below threshold, non-target is above threshold

For calculating the probability that a target is above a non-target, terms 1, 2, 3 are relevant, and these are added together in the above equation. Terms two and three are multiplied by ½ because only half of all outcomes will result in the target being above the non-target.

It turns out that the above equation corresponds to an ROC curve with three points: (r,a); (0,1); (1,0) with terms 1, 2, 3 of the equation corresponding to an area with a square and two triangles respectively. The validity of this linear interpolation for calculating such an area under the ROC curve is also shown by Davis and Goadrich (2006), and confirms the above reasoning for calculating ROC-AUC from a point value.

It is also possible to calculate an equivalent PR-AUC from a single point value. This is derived by constructing a precision recall graph with the same three corresponding points as the ROC curve but this time interpolation is not linear. The method for this interpolation is described in more detail by Davis and Goadrich (2006). This allows PR-AUC to be given for point values for the relative minimal distance or the active articulator algorithm shown in Table 3.2 although it should be remembered that comparisons across datasets are best made with ROC-AUC. According to the ROC-AUC evaluation measure, the binary feature distance algorithm performs best closely followed by the relative minimal difference algorithm. The active articulator algorithm performs worst, but doesn't miss any allophones.

3.2.3 Theoretical shortcomings in using binary features

In the next section the measure of binary feature distance is proposed for cross-language comparisons of phonetic transcripts. One theoretical shortcoming in using binary features is that they are more phonologically motivated than they are phonetically motivated. This may limit their suitability for cross-language comparisons. For example a Spanish sound written as [p]

in one transcript may have exactly the same voice-onset-time as an English sound written as [b] in another transcript (Williams, 1977). Even though these sounds have the same voicing, a direct comparison of the symbols suggests a difference of one binary feature; [voice]. This problem is partly due to the limited detail inherent in symbolic phonetic transcripts. The phonetic shortcomings of binary features may, in the future, be lessened by associating them with probability estimates. Probabilistic binary feature recognisers have shown promising performance for cross-language phone recognition (Siniscalchi et al., 2008).

Mielke (2009) has investigated a more phonetically motivated approach to phonetic similarity. This includes airflow, acoustic, laryngeal, and vocal tract measurements from three trained phoneticians who were native English speakers. The results are also compared with a phonological distance measure. This doesn't directly address the above issue of the limitation of symbolic transcriptions, and a combined universal similarity metric has yet to be defined. However, Mielke's study is the most comprehensive treatment of the subject of phonetic similarity to date, and should eventually lead to a more suitable similarity metric to be used in phonemic analysis.

Ideally phonetic features would be derived directly from the acoustic or articulatory data of the target language, but this is unfortunately rarely practical in survey scenarios. Sometimes for languages on the verge of extinction, phonetic transcripts are the only data that may be available (e.g. Chamicuro described in Section 3.4.2). Using binary features may be regarded as a simplistic model, but the arguments for using them are primarily pragmatic. As described in Section 3.1.1, the appeal is their simplicity for algorithmic implementation and their flexibility in representing speech sounds with multiple articulations. To this it could be added that they are an adequate representation when phonetic transcriptions are the only data available. In this thesis they have also been used in effective algorithms.

3.3 Dealing with sequences of sounds

The consonants in the TIMIT corpus are all associated with a single set of binary features. This includes affricates like [t͡ʃ] that are [-continuant] but otherwise share the same features as their fricative counterparts [ʃ]. However the vowels include diphthongs which are a sound sequence behaving as a single sound. This is known as a contour segment. Contour segments can be expressed by a sequence of binary feature sets. Before investigating sound sequences that behave as a single sound, it is useful to look at the more general problem of representing and comparing sound sequences.

The field of dialectometry (Kondrak, 2003) is relevant here because it is concerned with measuring the difference between sounds sequences in different dialects. A common approach (Nerbonne and Heeringa, 2010) is to measure the Levenshtein distance between word pairs i.e. the number of insertions, deletions and substitutions needed to convert one phone sequence to

the other. This is calculated with dynamic programming. For each comparison, this distance is usually normalised by the length of the longest phone sequence (Kondrak, 2003).

As an example, consider the two phone sequences:

Language A: [ʔ a^{55} ŋ̩^{21} k a^{55} l a^{55} m u^{33}]
Language B: [a ŋ k l a m u]

Language A has tone transcribed as superscript numbers that follow a phonetic convention of labelling 1 as a low tone and 5 as a high tone; the numerical equivalent of tone letters (Chao, 1930). Language B shows no change in tone so this is not transcribed.

When calculating the Levenshtein distance, only exact matches are allowed, e.g. [a^{55}] and [a] is not considered to be a match. In the above example, going from language A to B there are two deletions and four substitutions, giving a distance of six, which when divided by the length of the longer sequence gives a normalised distance of 67%. One problem with this approach is that it is not apparent from such a large distance how close the substituted phones were to each other. Another problem is that when only some phones match completely, the alignment can be incorrect. Figure 3.7 shows the dynamic programming working over a similar example, this time with an extra sound at the beginning of the word in Language B. The pause symbol (.) is taken from the extended IPA (Esling et al., 1999). If one more of the phones were slightly different e.g. if [k] in language A was retracted [k̠], then the alignment would fail, with the dynamic programming erroneously creating a mapping straight down the diagonal.

Gildea and Jurafsky (1996) defined a distance measure between two phones akin to the binary feature distance discussed above. Sequences of phones are then compared with dynamic programming. For this thesis it was decided that the binary feature approach of Gildea and Jurafsky would be adapted. In calculating the cumulative distance for phone sequences, Gildea and Jurafsky state “the cost of insertions and deletions was arbitrarily set at six, roughly one quarter the maximum possible substitution cost” (Gildea and Jurafsky, 1996). In the experiments for this thesis the dynamic programming, with uniform transition penalties, calculates the cumulative distance directly, without any further modification. This allows the cumulative distance to be given as the total number of binary feature edits. This can be normalised to give the average number of binary feature edits per phone (BFEPP).

Following the standard approach in dialectometry the normalisation should be calculated by dividing by the number of phones in the longest sequence. In the above example of language A and B the BFEPP measure gives a value of 3.3. Optionally a percentage score can be given by dividing BFEPP by the number of features in the binary features system. The same example comparisons of the two utterances is shown in Figure 3.8, this time with the dynamic programming executing on the BFEPP distances.

The scenario given above is outlined here because it is relevant to the cross-language issues raised in Chapter 4. However, the principal of comparing sequences can be applied to much

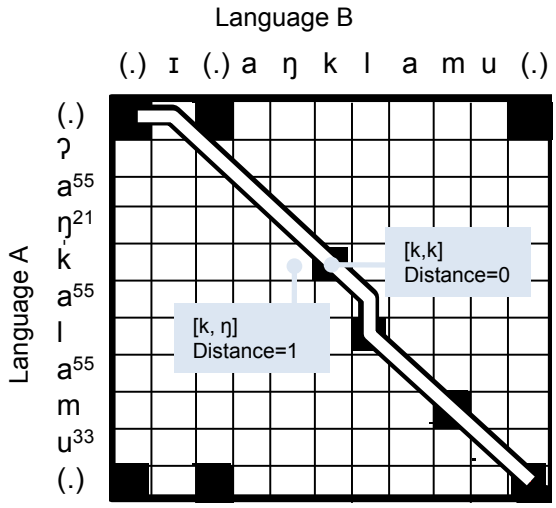

Figure 3.7: When calculating the Levenshtein distance, the distance between two phones is either an exact match or a non-match. Dynamic programming is used to calculate the distance between two sequences.

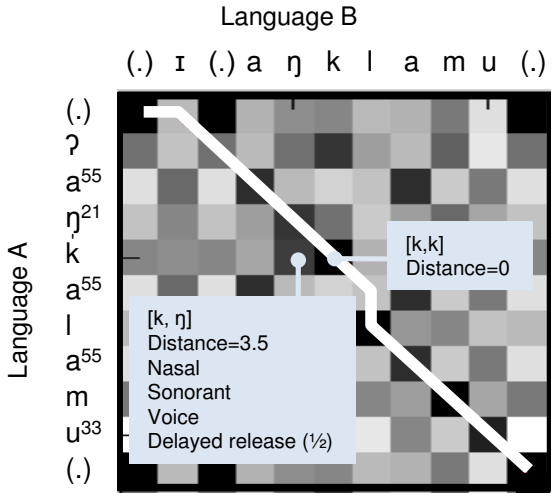

Figure 3.8: For calculating the binary feature edits per phone, the distance is the number of binary feature edits. Dynamic programming is used to calculate the distance between two sequences.

shorter sequences such as those sounds that can behave as a unit within a language such as diphthongs. It also allows many more types of sounds such as triphthongs, preglottalized sounds, and tone contours to be used in a phonetic similarity calculation. In fact in the example shown in Figure 3.8, tone contours are included e.g. [n̥^{21}]. The novelty of the BFEPP measure introduced in this thesis is that it works on both levels; sequences of phones and sequences of phone components.

The BFEPP approach allows a rough phonetic distance to be calculated between any utterance in any language. As would be expected with a universal phonetic distance measure, the applications of this tool are potentially very broad. In this thesis the measure is tested in the context of phonemic analysis but there are other areas where it could contribute. In the field of dialectometry, previous studies have found that the simple Levenshtein distance slightly outperformed feature based methods (Heeringa et al., 2006). More recently there is a renewed interest in the use of features for calculating phonetic distances, because feature-based methods tend to reduce the problem of multiple alignments (Nerbonne and Heeringa, 2010) i.e. result in alignments that are more robust to small phonetic changes. For example if [k] in language A was retracted [k̟] the alignment would be stable with BFEPP but it would fail with the Levenshtein distance. BFEPP could also contribute in the area of the assessment of speech disorders. When both vowels and consonants need to be taken into account in the analysis of a speech disorder, a common approach amongst practitioners is to count the percentage of phonemes correct (PPC) (Dollaghan et al., 1993). This is essentially equivalent to 100% minus the normalised Levenshtein distance, except in the original definition phoneme distortions are counted as an exact-match. Using BFEPP has the potential to address idiosyncrasies in the way phoneme distortions are treated, since other researchers such as Shriberg et al. (1997, p.712) treat phoneme distortions as a non-match (i.e. a substitution). Also, using the automated dynamic programming would speed up what is currently a manual process.

3.3.1 The relative minimal difference and sequences of sounds

The relative minimal difference algorithm was extended to handle contour segments. When checking whether a phone is phonetically in between two other sounds, if all components of the middle phone is phonetically in between all single component of the other sounds, the phone is counted as in between. This would mean the other two phones are not regarded as allophones. For example [ʔ͡m] and [p^{h}], would not be regarded as allophones if [p] also occurred in the language. This is because [p] is both phonetically between [ʔ] and [p^{h}] and also phonetically between [m] and [p^{h}].

3.3.2 The active articulator and sequences of sounds

The active articulator algorithm can also be extended to handle contour segments. A contour segment made up of other sound components takes on any active articulator that is used in the component sounds:

$$a = a_1 \cap a_2 \cap a_3 ...$$

For example, deriving the active articulator set for the sound [ʔ͡m]:

$[\widehat{ʔm}]_{AA} = [ʔ]_{AA} \cap [m]_{AA}$

$[\widehat{ʔm}]_{AA} = \emptyset \cap$ {labial, nasal}

$[\widehat{ʔm}]_{AA} =$ {labial, nasal}

3.4 Suitable corpora for experiments

With the algorithms extended to handle many different types of sounds, this section reports on potentially suitable corpora for evaluations.

3.4.1 Well-resourced languages

TIMIT (Garofolo et al., 1993) is a broadcast-quality corpus of US English and is one of the most extensive careful transcriptions ever produced. 1386 phonetically diverse sentences from the training subset of TIMIT were used in the experiments. Although there is some allophonic detail not included in the narrow transcriptions, about 25% of the phone set are environmentally conditioned allophones e.g. [ɦ] and [ʉ].

There are other corpora apart from TIMIT that have the potential to be suitable for experiments in machine-assisted phonemic analysis. These are shown in Table 3.3, with more phonological details in Table 3.4. The column in the latter table titled *data match* refers to the match between the language variety (e.g. dialect or accent) of the phonetic data, and the language variety of the phonological analysis. Since it is not possible to create a perfectly objective segmental phonetic transcription, it is appropriate to indicate which direction the bias is in; this is shown in the column *phonetic bias*. Phonetic bias can either be non-native where there is a risk of missing a contrast, or a native bias where there is a risk of missing allophonic detail. The extent of the phonetic bias is not indicated though all transcriptions, apart from the Bosnian data, are narrow phonetic and produced by expert phoneticians.

SCRIBE (Spoken Corpus of British English) (Huckvale et al., 1989)[2] was a pilot project to

[2]The original project involved many different authors. Huckvale later compiled and corrected the SCRIBE data providing a subset publicly available on the internet.

Corpus	Location	ISO693-3	Size	Audio alignment
TIMIT	USA	eng	6300 sentences	Phone-level
SCRIBE	UK	eng	280 paragraphs	Phone-level
Switchboard	USA	eng	4 hours	Syllable-level
OGI-MLT	Varies	Varies	1.7 hours	Phone-level
Kua-nsi	China	ykn	540 words	Utterance-level
Nisu	China	yiv	320 words	Utterance-level
Chamicuro	Peru	ccc	1000 words	No audio
Awing	Cameroon	azo	3000 words	No audio
Seri	Mexico	sei	300 words	Small amount
Bosnian	Bosnia	bos	3.8 hours	Utterance-level

Table 3.3: Corpora for testing machine-assisted phonemic analysis

create a British corpus similar to TIMIT. A small number of sentences were given a narrow transcription. This included passages, sentences and a small amount of free speech. This narrow transcription uses a SAMPA (Wells et al., 1992) variant for the transcription. A description of the ASCII-based diacritic symbols can be found in Hieronymus et al. (1990).[3]

Switchboard is a corpus of telephone conversations between strangers in US English. The switchboard transcription project (Greenberg et al., 1996) has resulted in narrow transcripts including a different set of allophones from TIMIT e.g. [fi] "has been omitted as it is generally a contextually predictable variant" but extra diacritics are used to indicate other allophones. In common with TIMIT, the phonetic transcription has been completed by native speakers, which is indicated in the phonetic bias column.

OGI-MLT is a multilingual telephone corpus, developed at what is now the Centre for Spoken Language Understanding at the Oregon Health and Science University (Muthusamy et al., 1992). A small part of the corpus has been narrowly transcribed. The transcriptions are from spontaneous monologues in six languages: English, German, English, Hindi, Japanese, Mandarin and Spanish and do not include word boundaries. A plethora of ASCII-based diacritics are used in the transcriptions indicating a large amount of phonetic detail; however many common allophones (such as differences in aspiration of stops in English) are not distinguished. For most of the languages, the language variety was not documented, and many of these accents differ from the only published phonemic analysis of the standard variety. That is why the data match is described as fair.

[3] This paper was once considered lost but in the process of examining this thesis the internal examiner discovered a hard copy of the paper (to the delight of those wishing to make full use of the SCRIBE narrow transcriptions as well as the paper's author)

Corpus	Phonemes	Allophones	Data match	Phonetic bias	Analysis
TIMIT	39	25%	Good	Native	Mature
SCRIBE	44	20-30%	Good	Native	Mature
Switchboard	39	24%	Good	Native	Mature
OGI-MLT	Varies	Varies	Fair	Native	Mature
Kua-nsi	59	14%	Fair	Non-native	Developing
Nisu	47	11%	Good	Native	Developing
Chamicuro	29	20-30%	Good	Non-native	Quite mature
Awing	35	20%	Fair	Non-native	Quite mature
Seri	24	14-50%	Good	Native	Mature
Bosnian	30	0%	N/A	Native	Mature

Table 3.4: Phonological details of corpora

3.4.2 Under-resourced languages

Kua-nsi (Castro et al., 2010) is part of the Tibeto-Burman language family. It is a tone language, i.e. changes in pitch can distinguish words. It is in the process of becoming a written language with an orthography being developed. Therefore the analysis performed to date is still not completely mature and there may be small revisions in the future. The phonemic analysis (Lehonkoski et al., 2010) was based on a variety of the language spoken in the village of San'gezhuang, Heqing county, Yunnan. Comprehensive phonetic data though is only available from other villages. The nearest village Hedong is 3km south of San'gezhuang but across a steep ravine. This means the communities do have some separation between them, and there is evidence from current surveys that that the accents are different[4] with possibly slight differences in phonology. Therefore the data match is described as fair. The phonetic transcript was produced by a non-native speaker before the attempt at a phonemic analysis, as indicated in the phonetic bias column.

Nisu (Yang, 2009) is also a tone language from the Tibeto-Burman language family. Nisu has a traditional logographical orthography used for religious purposes but literacy is limited to a few shaman. A phonemic analysis has been completed on the northern variety spoken in Laochang, Xinping county, Yunnan. A phonetic transcript was then produced with knowledge of that phonemic analysis[5]. That is why the phonetic bias is described as native.

Chamicuro (Parker, 2010) spoken in Peru is a critically endangered language, and might be extinct already. It is part of the Arawakan family. A large 1000 word list has been published with a narrow transcription in Americanist notation, and this includes most of the words from

[4]Crook (2011), personal communication
[5]Yang (2011), personal communication

the Swadesh 100 list. The main fieldworker reports that no audio has been known to have been collected.

Awing (van den Berg, 2009; Alomofor and Anderson, 2005) is a tone language spoken in Cameroon and is part of the Niger-Congo family. It has been narrowly transcribed and is likely to need some consistency checking on the (unpublished) IPA transcripts before the data can be used in computational experiments.

Seri is a language isolate spoken in Mexico. It is used as a case study for phonological analysis; and there is a high maturity to the analysis (Marlett, 2005). Most narrow transcripts that exist are for the purpose of illustrating particular surface forms, and a restricted set of allophones are included for teaching particular phonological phenomenon. These are derived from a good knowledge of the underlying forms which is why the phonetic bias is described as native.

The Bosnian dataset (Kurtic et al., 2012) does not have any narrow transcriptions but is relevant in the area of forced alignment. Each speaker was recorded separately so there are four channels of audio. The original manual alignment is at the utterance-level. Phone-level alignment was created later as part of the work described in Chapter 4. Bosnian, like English, is part of the Indo-European language family.

As described in Section 1.4 the primary target application scenario for machine-assisted phonemic analysis is the use of survey data that typically has a non-native bias which should be reflected in the evaluation. For accurately evaluating results the phonological analysis should be mature. Generally, it is difficult to find a dataset where both the phonetic bias is non-native and the phonological analysis is mature. This can be for a variety of reasons, e.g. the analysis could have been quite informal at first, with any subsequent data collection affected by native phonetic bias. Also, over the years it takes for a phonological analysis to become mature, the original survey data may have been lost. Although in the short-term digital solutions have arguably made the situation worse, it is hoped that the archiving of primary documentation will become more routine for linguists (Bird and Simons, 2003). This will mean more suitable data for testing machine-assisted phonemic analysis is likely to be available in the future. In the meantime the evaluation is performed on most suitable corpora that can be found. The Kua-nsi corpus was chosen for the evaluation alongside TIMIT, since Kua-nsi had a non-native phonetic bias, and audio material was available to allow for any reinterpretations of the transcriptions.

3.4.3 The algorithms applied to Kua-nsi data (Experiment 3D)

The phonetic similarity algorithms were applied to the Kua-nsi language data. The data is from Castro et al. (2010) and the ground truth of allophone pairs is included in Appendix B. The results are shown in Table 3.5 and 3.6. Again the focus was on consonants to make it comparable to previous experiments but this time contour segments such as [ʔn͡] were included.

	Rel. min. difference	Articulator	Combined
Hits	5	6	5
Misses	1	0	1
False alarms	170	415	114
Correctly rejected	644	399	700
False alarm rate	0.209	0.510	0.140
Recall	0.83	1.00	0.83
Precision	0.0286	0.0143	0.0420

Table 3.5: Results of both the allophone detectors on Kua-nsi

Algorithm applied to Kua-nsi	ROC-AUC	PR-AUC
Binary feature distance (BFEPP)	87.0%	4.8%
Relative minimal difference	81.2%	2.7%
Active articulator	74.5%	1.4%

Table 3.6: Area under the ROC and PR curves for the different algorithms on the Kua-nsi corpus

The different algorithms show the same ranking of performance when compared with the TIMIT results. Again, the active articulator algorithm does not miss any allophones but has many false alarms. The binary feature distance measure is the most successful.

3.4.4 The algorithms applied to a French phone set (Experiment 3E)

In previous studies it appears that the relative minimal difference algorithm was not tested on its own, so it is not possible to make a direct comparison with the results in this thesis. However, results from Peperkamp et al. (2006) indicate that the relative minimal difference algorithm has a good precision and recall. For example it reduces the false alarms of the main complementary distribution algorithm from 129 to 8 and yet manages to preserve all the hits for the 7 allophones detected. Is this due to an easier dataset or are the multi-valued features superior to binary features? To find out, the algorithms described in this chapter were evaluated with Peperkamp's French data; specifically 21 consonants plus 9 allophones. Results are shown in Table 3.7. The result for the minimal distant filter show an improvement when compared to English (TIMIT) and Kua-nsi. The high PR-AUC results mean there are less false alarms and alongside the other results it indicates that the French dataset is less challenging. This does not rule out a different performance between the two features sets, but it does show that the dataset is a significant reason for the difference. Again, the ranking of the algorithms is the same as the previous experiments, with BFEPP performing best.

Algorithm applied to French	ROC-AUC	PR-AUC
Binary feature distance (BFEPP)	99.0%	52.6%
Relative minimal difference	94.7%	16.7%
Active articulator	74.9%	4.0%

Table 3.7: Area under the ROC and PR curves for the different algorithms on the French data

3.5 Conclusions

The relative minimal difference algorithm introduced by Peperkamp et al. (2006) and adapted in this chapter to work on all languages, has been shown to help detect allophones among the consonants in US English (TIMIT Experiment 3A) and Kua-nsi (Experiment 3D). The data used in these experiments is more faithful to the acoustic signals than in previous experiments.

With the introduction of ROC-AUC as the primary evaluation measure, all the algorithms evaluated in this chapter are shown to perform better than chance (i.e. ROC-AUC $>$ 50%) on all the languages tested.

The new active articulator algorithm, shows a lower ROC-AUC performance than the relative minimal difference algorithm, but has a higher recall of allophone pairs; consistently at 100% (Experiment 3B). Although there are many false alarms, no allophones are missed in English, Kua-nsi, French or Japanese. Combining the active articulator algorithm and the relative minimal difference algorithm is shown to improve the PR-AUC measure.

Most of the algorithms produce a large number of false alarms which is reflected in the relatively low PR-AUC values. Previous studies on different language data such as a French dataset appear to show a smaller number of false alarms. Experiment 3E tested the algorithms on the French data giving much improved PR-AUC values with a reduced number of false alarms. This indicates that the data used in this thesis is more challenging than in previous studies, rather than there being a problem with the algorithms or the feature system used.

The new binary feature distance algorithm (BFEPP, Experiment 3C) is shown to perform the best with highest ROC-AUC values on all the languages tested (Experiment 3A, 3D, 3E).

Phonetic similarity is not sufficient on its own for determining all the allophones in a language. And it is not used by linguists on its own for a phonemic analysis. This is one reason why the results do not show a performance closer to 100%. The experiments were completed because it is important to understand the standalone contribution of phonetic similarity to a phonemic analysis (see Section 7.1.2 in the final chapter). The more interesting result is the comparative performance of the algorithms, which the evaluation framework presented in this chapter allows by providing a quantitative measure.

3.6 Chapter summary

In a phonemic analysis, relying on some notion of phonetic similarity is sometimes implicit, but it is always important. In this chapter different phonetic distance heuristics are evaluated quantitatively for their effectiveness in detecting allophones.

Binary features were used for the phonetic representation due to their simplicity for algorithmic implementation and their flexibility in representing speech sounds with multiple articulations. The phonetic shortcomings of binary features may in the future be lessened by associating them with probability estimates.

Three different phonetic distance algorithms were evaluated; relative minimal difference, the active articulator, and the binary feature distance. The relative minimal difference algorithm was adapted from earlier work by Peperkamp et al. (2006) with the other two algorithms proposed in this thesis. Each algorithm has been generalised to work with contour segments e.g. the binary feature distance is generalised to BFEPP (binary feature edits per phone). The evaluation was performed on the best datasets available: the TIMIT and Kua-nsi corpora. To perform the evaluation a number of information retrieval evaluation metrics were used; the most important being ROC-AUC which allowed each algorithm to be compared via a single performance figure.

All algorithms performed better than chance, the binary feature distance algorithm performed the best, followed by the relative minimal difference. The active articulator algorithm performed the worst but had the advantage of never missing an allophone pair. There were a high number of false alarms in most of the experiments to detect allophones. Results suggests that this is due to the use of challenging datasets that are more faithful to the acoustic signal than in previous studies.

The results show that a phonetic distance algorithm such as BFEPP can contribute to the phonetic similarity judgement procedure to assist in a phonemic analysis.

The significant original contributions of this chapter are as follows:

- Three phonetic similarity algorithms were evaluated on the TIMIT and Kua-nsi corpora
- Statistical measures have been applied to quantitatively evaluate phonemic analysis
- The data used in the experiments is more phonetically accurate than previous studies
- The active articulator algorithm was introduced for predicting phonemically distinct phones
- The BFEPP algorithm was introduced as the best performing phonetic similarity measure

Parts of this chapter have been published in Kempton and Moore (2009) and Kempton et al. (2011).

Chapter 4

Phone recognition and alignment

Creating an impressionistic narrow phonetic transcription is one of the earliest stages in a phonemic analysis, and it is a highly skilled and lengthy task. Alignment of the phonetic transcription with the audio is a less common task but can be used to investigate particular phonological phenomenon. In this chapter both automatic phone recognition and automatic phone alignment are investigated. The structure of this chapter is as follows. Section 4.1 explains the challenge of the task, and defines evaluation measures. Section 4.2 outlines experiments on cross-language phone recognition on the TIMIT corpus. Section 4.3 introduces cross-language forced alignment, with experiments both on the TIMIT and Bosnian corpora. Section 4.4 gives the conclusions.

4.1 The challenge: minimum knowledge of the language

Cross-language transfer is a term used to describe the recognition of a target language without using any training data from that language. The technique is particularly useful for languages lacking labelled data and other linguistic resources (Schultz and Kirchhoff, 2006). Most work on cross-language transfer has assumed a certain amount of target language knowledge such as pronunciation dictionaries, or at the very least a phoneme inventory (Schultz and Waibel, 1998; Siniscalchi et al., 2008). However, for many languages, the phoneme inventory is not known. As explained in Section 1.2, deriving an inventory of phonemes (contrastive sounds) from phones (sounds with an unspecified contrastive status) is a non-trivial task.

There appears to have been only one previous paper that has explicitly addressed the problem of cross-language phone recognition with minimal target knowledge. Walker et al. (2003), attempting to build a universal phone recogniser, included an experiment where there was no knowledge of the target language phoneme inventory. This resulted in a phone recognition accuracy of 13%. When knowledge of the inventory was included, the accuracy doubled. Clearly, cross-language phone recognition is already a challenging task, but even more so when there is

no knowledge of the phoneme inventory. This is because there are more phones to distinguish. Ideally every possible phonetic contrast that might occur in a language needs to be detected.

In this chapter, phone recognisers developed by Schwarz et al. (2009), for Czech, Hungarian and Russian, were evaluated on a target language where the phoneme inventory is not assumed to be known. The rationale for using phone-based models is given in Section 2.4; these particular phone recognisers were chosen because they are state-of-the-art and publicly available. This type of evaluation is characterised by the ground truth containing fine phonetic detail. To improve the evaluation, the BFEPP measure (binary feature edits per phone, introduced in Section 3.3) was used alongside PER (Phone Error Rate). The BFEPP measure led to a novel adaptation of a ROVER method to make it phone-based rather than word-based. The measure also led to the new technique of automatic phone set mapping for cross-language forced alignment.

4.1.1 An illustration from an unwritten language

Kua-nsi is a Tibeto-Burman language spoken in the Yunnan province of China that has no writing system of its own although an orthography is currently being developed. More information can be found in Section 3.4.2. Initial documentation of the language has been completed by Castro et al. (2010). The description of the language includes a list of over 500 words with impressionistic phonetic transcriptions representing more than 100 sounds including the tones. Audio recordings of the words were obtained from the authors. A sample of the wordlist is shown in Table 4.1. This was an early survey so there was little knowledge of which sounds contrasted with each other i.e. the phoneme inventory was not known. As discussed in Section 1.2, the linguists have to be as objective as possible transcribing every detail heard that might turn out to be significant. For example, the nasalised vowel, and the aspirated africate would not be significant in English but might be significant in this language. As well as identifying the sounds, the process of transcription is also a process of choosing a particular segmentation e.g. for the second syllable of the word *rain* the linguist decided on the transcription [huã^{55}] rather than [ʍã^{55}].

The Kua-nsi corpus is ideal for evaluating cross-language phone recognition with minimum target knowledge. However, the audio often contains a word said in Chinese to elicit the response; this is spoken at some distance from the microphone but is still partially audible. Before a full scale evaluation the audio needs to be trimmed so that it matches the transcriptions.

A recording of the first word in the list, the Kua-nsi word for *sky* in the Hedong dialect, can be used to illustrate some of the principles of this study. In this example a Czech recogniser was used for the cross-language phone recognition:

Kua-nsi transcription: [ʔ a^{55} ŋ̩^{21} k a^{55} l a^{55} m u^{33}]
Czech recogniser: [a ŋ k l a m u]

English	Hedong
sky	[ʔa^{55}.ŋ̩^{21}.ka^{55}.la^{55}.mu^{33}]
sun	[u^{55}.ts^{h}u^{55}]
moon	[hõ^{33}.bu^{33}]
star	[u^{55}.tɕua^{55}]
cloud	[tsɿ^{55}]
wind	[mi̩^{21}.hi^{55}]
rain	[ʔũ^{21}.huã^{55}]
lightning	[ʔn̩^{21}.bia^{21}.bia^{21}]
thunder	[ʔn̩^{21}.gɯ̩^{21}.t^{h}ua^{33}]
rainbow	[ʔu^{55}.ju^{21}.sua^{55}.zɿ^{55}]

Table 4.1: The start of a wordlist for Kua-nsi spoken in Hedong

It can now be seen that the utterances introduced previously in Section 3.3 were actually from the Kua-nsi language and the equivalent output from a Czech phone recogniser.

Phone recogniser labels were converted from a SAMPA variant to IPA Unicode, using the SAMPA specification (Wells et al., 1992) and the documented phonology of the relevant language (Esling et al., 1999; Maddieson, 1984; Mielke, 2008). The mappings are documented in Appendix C.

The normalised Levenshtein distance (as referred to in Section 3.3) can now be interpreted as how accurate the recogniser is performing. In fact the phone error rate (PER) is exactly the same as the normalised Levenshtein distance. So the accuracy of the recogniser can be given as 67% PER (two deletions and four substitutions).

4.1.2 Evaluation measures; PER and BFEPP

The standard tool to calculate *word* error rate in speech recognition – SCLITE – (NIST, 2009) can be used to calculate the PER. The tool uses dynamic programming to align the sequences and calculate the cumulative distance of insertions, deletions and substitutions (i.e. the Levenshtein distance). This is then normalised by the length of the reference transcription.

The use of dynamic programming to align two sequences was extended to multiple sequences for combining different speech recognisers. This is called ROVER (Fiscus, 1997) and is explored further in Section 4.2.3.

The BFEPP (binary feature edits per phone) distance measure introduced in Section 3.3 can also be interpreted as an evaluation measure. The distance is measured between the recogniser output and the ground truth reference sequence. The normalisation is calculated by dividing by the number of phones in the reference sequence. In this case the E in BFEPP can also be interpreted as the binary feature *errors* per phone. In the example given in Section 4.1.1, the

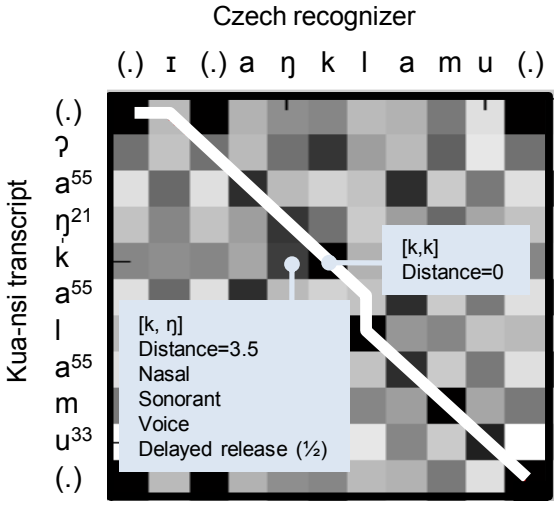

Figure 4.1: The distance between two phones is the number of binary feature edits. The cumulative phonetic distance between two phone sequences is calculated with dynamic programming.

BFEPP measure gives a value of 3.3. The dynamic programming of a larger part of the utterance is shown in Figure 4.1. The first vowel picked up by the Czech recogniser was the interviewer saying a word in Chinese. The limitation of the PER is that it is such a strict measure. BFEPP is arguably a better measure than PER because it more effectively expresses when there is a close match. PER is a very strict measure and will penalise the Czech recogniser for recognising [a^{55}] as [a] which is the best the recogniser could have done. It is the same penalisation as if it had recognised this vowel as a consonant. That is why the PER seems so high.

4.2 Cross-language phone recognition (Experiment 4A)

4.2.1 Experimental set-up

Since there is currently a lack of suitable data available in unwritten languages like Kua-nsi, a more conventional dataset was used for the evaluation in this thesis. The TIMIT corpus (Garofolo et al., 1993) was chosen because it is one of the only corpora that contain a large number of manually annotated allophones. The evaluation was conducted on the core test portion of 192 utterances; this portion was used to give a quick turnover of experiments and is a sufficient size for the statistical tests. A 53-phone set was used to include a maximum set of allophones, effectively meaning the phoneme inventory was not known, or was at least not well defined. In comparison, most studies combine the allophones to create a simpler 39-phone set (Schwarz, 2009, p.16) which is equivalent to the phoneme inventory of many US English dialects. The only combination of sounds in the work reported here was the merging of stop closures with

corresponding releases, and the merging of the epenthetic silence with adjacent phones.

The phone recognisers are artificial neural network (ANN) based with a left context right context (LC-RC) architecture (Schwarz, 2009, Ch.5). The ANN produces posterior probability outputs corresponding to three states for each phoneme (Schwarz, 2009, p.42) every 10ms. These phone posteriors are then processed by a Viterbi decoder to produce a sequence of phones. The ANN takes as input 310ms of data which means that context dependency is implicit in the system, although it is not explicit context dependency as in a triphone model where there are separate models trained for different contexts (Schwarz, 2009, p.39). In terms of language modelling it is an unconstrained phone recogniser because there is no language model used in the system.

Performance of these phone recognisers is close to state-of-the-art as can be seen when the training and test languages match. The TIMIT recogniser performs at 24.4% PER (Schwarz, 2009, p.42), although with some tuning this can be improved to 21.5% PER (Schwarz, 2009, p.46). The other recognisers average at 35.1% PER (Matějka et al., 2005).

The Czech, Hungarian, and Russian phone recognisers process telephone bandwidth audio, so the TIMIT test data was downsampled. This conversion from a sampling frequency of 16kHz to 8kHz was achieved with the Sox tool, version 14.3.0 (Bagwell et al., 2009), with default parameters except for the dither option, which is turned off to allow others to generate exactly the same dataset. A closer match between training and test data channel characteristics would have been achieved by using a band pass filter, or better still, a version of TIMIT passed through a telephone network: NTIMIT. Previous cross-channel experiments on TIMIT and NTIMIT indicate accuracy would be slightly better on telephone channel test data (Schwarz, 2009, p.57). However, the setup described in this paper is closer to the application scenario, where the field linguists will often only have access to simple signal processing tools.

4.2.2 Direct cross-language phone recognition

Cross-language phone recognition was first performed directly. The mean error rates for the utterances are shown in Table 4.2. For the other chapters in this thesis the evaluation measures show greater numbers for a better performance. Since it is different in this current chapter with errors as the primary measure, the best performing recogniser is written in bold for clarity. The Czech and Hungarian recognisers, but not the Russian recogniser, show a greater accuracy than the 87% PER equivalent in Walker et al. (2003) which also made minimal assumptions about the target language. However, this previous study was conducted on *conversational* telephone speech in a *different* language, so the comparison should be interpreted cautiously. Note that the results appear very poor when PER is used. As discussed above this is because of the strictness of the measure.

Recogniser	PER	BFEPP
Czech	**73.5%**	**3.19**
Hungarian	80.7%	3.32
Russian	90.9%	3.74
ROVER vote, Czech breaks ties	77.7%	3.22
- without phonetic-align	76.9%	3.29
ROVER maximum score	79.8%	3.34
- without normalisation	83.6%	3.45

Table 4.2: Cross-language phone recognition on TIMIT; showing mean error rates for the utterances (Experiment 4A)

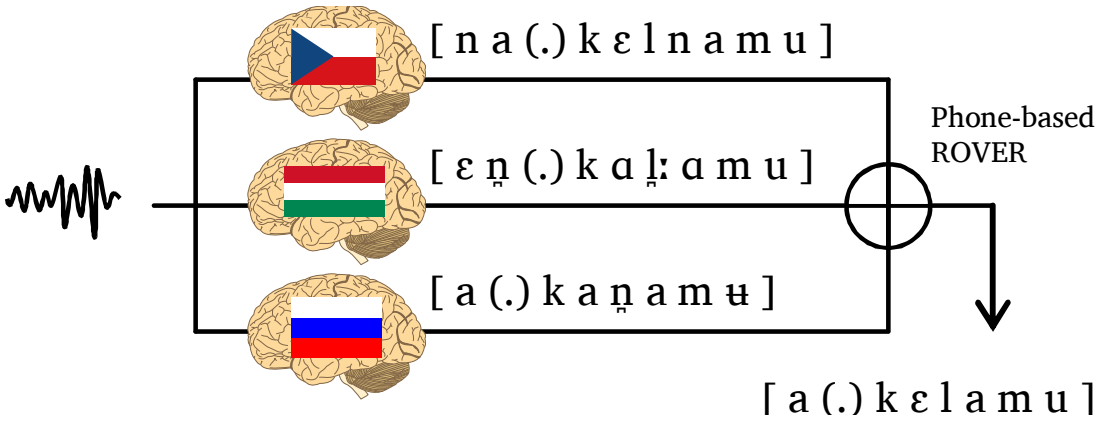

Figure 4.2: Phone recognition of the second utterance of Kua-nsi word [ʔa^{55}ŋ̩^{21}ka^{55}la^{55}mu^{33}] showing Czech, Hungarian, Russian and ROVER results

4.2.3 Phone-based ROVER

ROVER (Recogniser Output Voting Error Reduction) (Fiscus, 1997) is an algorithm that combines the output from multiple speech recognisers to reduce the overall word error rate. The algorithm works by aligning the output of the different recognisers and combining these results by conducting a vote for each word. ROVER evolved out of the SCLITE tool for evaluating speech recognisers, and was able to give a significant reduction in the word error rate. This is due to the underlying speech recognisers exhibiting error patterns that are not strongly correlated.

Here, the technique is applied to phones instead of words. The purpose is to take advantage of the alignment algorithm for multiple phone recognisers. Once the alignment has been completed, there are a number of options on how to combine the results to attempt greater accuracy and phonetic detail.

An example result on the second Kua-nsi utterance for the word *sky* is shown in Figure 4.2.

The phone-based ROVER system was implemented using the SRILM toolkit's `nbest-lattice` program (Stolcke, 2002) (SRILM version 1.5.8). Phonetic alignment, similar to that described in Section 4.1.2, was performed through a specially created dictionary file to give Hamming distances between phones. The normal purpose of using a dictionary file with the `nbest-lattice` program is to give Levenshtein distances between words. This is illustrated with two example entries in a dictionary:

```
Darnall  d ɑː n ə l
Donald   d ɒ n ə l d
```

The Levenshtein distance is 2 (one substitution and one insertion). It is possible to construct a dictionary to only allow substitutions by using placeholder words, thereby giving the Hamming distance rather than the Levenshtein distance. This is illustrated with the following two entries in such a dictionary, a dictionary that is a look-up table for phones rather than words:

```
ɛː   + + syllabic , - - stress , + + long , - - consonantal , + + sonorant
     , + + continuant , + - delayed_release , ...
ɦ    - - syllabic , - - stress , - - long , - - consonantal , - - sonorant
     , + + continuant , + + delayed_release , ...
```

The feature names and commas provide a human readable indication of which feature is being described, but their primary purpose is as placeholders to force a Hamming distance calculation. Feature values make use of two sign symbols to allow for unspecified values (indicated by "+ -"). The number of substitutions between the above two dictionary file fragments is equivalent to a difference of 3.5 binary features. Adding extra characters in the dictionary required a corresponding modification to the insertion and deletion penalty in the software[1].

The results of the different ROVER variations are shown in Table 4.2. The simple vote recogniser is very similar to the Nist1 vote in the original ROVER study (Fiscus, 1997). The maximum score voting recogniser is similar to the Nist3 vote (with alpha set to zero). These voting schemes are accompanied by two different variants. The simple vote without the phonetic alignment only allows exact matches, and the maximum score without normalisation are the raw scores without a simple normalisation of the means. A one-way repeated-measures ANOVA was used to test for statistical significance. The factor *recogniser* had a significant effect for both the PER measure and BFEPP measure (both $p < 0.001$). In comparing the PER and BFEPP measure, the average correlation across the seven recognisers was 0.57 and PER showed less variance. This may be simply due to the fact PER and BFEPP are measuring different types of errors.

The results appear disappointing, with none of the scores improving on the best compo-

[1] The minLength variable was divided by 4 in the file VocabDistance.cc

nent recogniser. The simple voting method performed best but this had a bias towards Czech. Surprisingly there was only a small difference in using phonetic alignment, and both measures disagree on whether this is beneficial.

Visual inspection suggested that phonetic alignment did improve results, but that all methods produced many alignment errors. An attempt was made to manually tune the insertion-deletion penalty, and it was found that there was an apparent optimum setting.[2] This did reduce the alignment errors by approximately half, but there was only a small improvement in recognition rates and they still did not surpass the Czech performance.

Future work with phone-based ROVER

Why is it that phone-based ROVER does not appear to show the same success as word-based ROVER? Word-based ROVER usually constructs correct alignments of recogniser outputs (Fiscus, 1997). This is ensured by having an adequate level of accuracy for each component recogniser. The low accuracy for phone-based recognisers cause more misalignments for phone-based ROVER, which is apparent in this experiment. Alignment accuracy is expected to improve as phone recognition accuracy improves in the future. Good ROVER results require accurate alignments, but independent recognizer components are also important. For ROVER to work well, error patterns of component recogniser outputs should not show a strong correlation. If the simple vote algorithm (where Czech breaks ties) performs worse than the Czech recogniser, it suggests that the other recognisers wrongly outvote the Czech recogniser many times. This indicates a problem with correlated errors and a bias towards the same errors in different recognisers.

Further diagnosis of recogniser errors would be helpful. Unfortunately there is an incompatibility with some of the functions of SCLITE and Unicode, e.g. so that confusion matrices cannot be easily produced. If this isn't fixed in subsequent versions, future work could involve re-encoding phone sets so that a better breakdown of the errors can be produced.

Different options for combining phones from different recognisers may help. These could be considered before adding further recognisers to the system. One of the limitations with simple voting occurs when the target language contains a sound that rarely occurs in other languages. Even if this exists in the phone set of one of the recognisers, it will be outvoted. An analysis of the recogniser phone sets could be conducted to give each sound an equal priority. Another option is to use feature-based voting.

The Kua-nsi corpus contains each word repeated three times. A ROVER approach could take advantage of this by combining the repeated utterances. As a proof of concept, the example audio described in Section 4.1.1 was processed through the different recognisers and the ROVER

[2]The insertion-deletion penalty is effectively the minLength variable and the optimum setting occurred when it was reduced by a further third (i.e. minLength was divided by 12)

simple voting combination. The three repeated utterances were then combined using the same ROVER method:

Utterance 1:	[(.) a ŋ k a l a m u (.)]
Utterance 2:	[(.) a (.) k ɛ l a m u (.)]
Utterance 3:	[(.) ɛ ŋ k a l ɛ m u (.)]
Result:	[(.) a ŋ k a l a m u (.)]

The result is more accurate than the Czech recogniser result in Section 4.1.1; a vowel is now included between [k] and [l]. This is not reflected by the PER which is still at 67% (a deletion is replaced with a substitution) but it is reflected by a drop in BFEPP from 3.3 to 2.4.

4.3 Cross-language forced alignment (Experiment 4B)

The analysis of under-resourced languages often requires the alignment of phonetic transcripts with audio. This can facilitate automatic acoustic analysis of particular phones e.g. formant frequency plots for vowels or a voice onset time histogram for stops. Automatic alignment is very different from the problem described in the above experiments, because the phonetic transcription is provided by the linguist. Alignment of a transcript and audio can be achieved with the well known process of *forced alignment* (see Jurafsky and Martin (2008, Ch.9) for a helpful explanation). However, when the amount of data is very small it makes it difficult to train or adapt acoustic models, especially if the phoneme inventory is not known. Usually forced alignment is performed on a *phonemic* transcription derived from the original text and a pronunciation dictionary. When the phonemic inventory is not known the transcription is *phonetic* and there are many more possible sounds which would require much more training data.

To address this problem, the concept of cross-language forced alignment is introduced. This is similar to cross-language phone recognition, except that the phone transcript is already provided. This transcript, which uses the recogniser phone set, is derived from the original language transcript via a suitable transformation. Forced alignment gives the timings which are then used for the transcript labels in the original language.

In this experiment the transformation consisted of mapping each phone in the original transcript to the closest phone in the recognizer phone set automatically, using the BFEPP distance measure. The same phone recognisers used in the recognition experiments were used in the forced alignment experiment. This technique has some similarities to the work of van Niekerk and Barnard (2009), who mapped phones from an under-resourced language to broad phonetic class labels derived from TIMIT. The difference is that the mapping method described in this thesis is language-universal and automatic. Also the work here does not use multiple iterations to

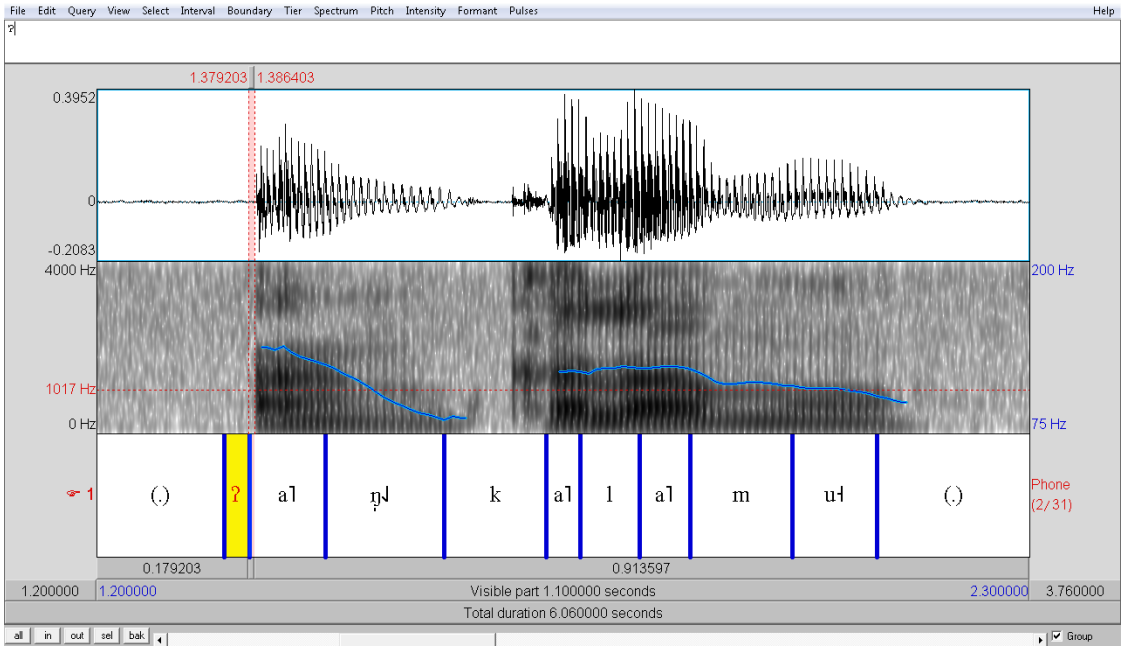

Figure 4.3: Cross-language forced alignment on Kua-nsi using Czech phone models (tone letters are used instead of tone numbers)

refine such alignments, which has the potential to improve results. In fact Van Niekerk has run experiments showing that multiple iterations reduce the errors to half of what they were[3]. The reason that multiple iterations where not used in this experiment was to prioritise pure cross-language baseline results, and also it was non-trivial to adapt the particular phone recognisers used.

The performance of each forced alignment was evaluated with software provided by Hosom (2009). This gives the standard forced alignment evaluation error; the proportion of boundaries placed more than 20ms away from the manually labelled boundary.

An example forced alignment result on the Kua-nsi utterance for the word *sky* is shown in Figure 4.3. The linguist's transcript uses tone letters (Chao, 1930), partly because the Praat software cannot handle the superscript tone numbers. The phones were first automatically mapped to the Czech phone set using the BFEPP measure. The closest phone to [ʔ] in the Czech phone set is [p] and this was used for the first sound in the word. The endpoint occurs slightly too early compared to the waveform, but is within the 20ms tolerance. The overall alignment appears to have a high accuracy.

As described earlier in the recognition experiment, the Kua-nsi dataset is not yet ready for full scale evaluation, so the TIMIT corpus was used instead. The results for single recognisers

[3]Van Niekerk (2011), personal communication

Recogniser	20ms Error
Czech	39.0%
Hungarian	42.7%
Russian	43.4%
Mean combination	38.8%
Median combination	**35.9**%

Table 4.3: Cross-language forced alignment on TIMIT (Experiment 4B)

are shown in Table 4.3. The performance ranking for the three languages is similar to the recognition task.

Following the forced alignment experiments based on single recogniser; two combination methods were investigated. The first combination method involved taking the mean time value of the three single recogniser boundary positions to create a new hypothesised boundary. The other combination method involved using the median of the three single recogniser boundaries. Table 4.3 shows that the median method was the most successful, with a small improvement over the best single recogniser. A one-way repeated-measures ANOVA was used to test for statistical significance. The Greenhouse-Geisser correction was used because the variances of the error values were not equal when comparing recognisers. The factor *recogniser* was shown to have a significant effect ($p < 0.001$). Using pairwise comparisons with the Bonferroni correction, there was a significant difference between each recogniser (all $p < 0.001$) except Czech compared to the mean combination ($p > 0.99$) and Hungarian compared to Russian ($p > 0.99$).

For many acoustic analysis tasks, e.g. vowel formant plots, the accuracy rates of all the recognisers are high enough to be useful to field linguists. If a linguist requires boundaries to be within 20ms, then less than half of boundaries need correcting. For under-resourced languages the alignment process is usually done by hand, so this approach can significantly reduce the time needed i.e. major errors can be quickly identified visually.

The results can be compared with other techniques for forced alignment. The performance is much higher when the challenge of cross-language recognition is removed. The best phone forced alignment system carefully trained and optimised for English, resulted in a 20ms error rate of 7% equivalent to the intra-transcriber agreement of expert human phoneticians (Hosom, 2009). A tool that has been designed for linguists to use called *EasyAlign* (Goldman, 2011) gives less accurate results: when optimised for English or French it gives an average 20ms error rate of 22%. Conversational data is more challenging, for example Kurtic et al. (2012) show a 20ms error of 35% on British English meeting data. A general purpose acoustic phonetic tool *Speech Analyzer* (SIL, 2007) has experimental functionality which is believed to be the only language universal forced-alignment feature available to linguists. Based on initial experiments, this resulted in an estimated 20ms error rate of over 90%.

TIMIT phone set	Closest Russian phone chosen	BFEPP
p	p	0
b	b	0
m	m	0
m̩	m	1
f	f	0
v	v	0
θ	s̪	1
ð	z̪	1
ɾ̃	r̪	4

Table 4.4: Part of an automatic phone set mapping for expressing the TIMIT phone set with Russian phones

A paper by Peddinti and Prahallad (2011) shows on their data that acoustic phonetic refinements to an original rough forced alignment can dramatically improve results. These acoustic phonetic refinements exploit phone-class specific sub-band energy events. Their best result gave a 20ms error rate of 11% on the Telugu language. The authors also confirm that a flat-start forced alignment (where HMM models are given no original training) results in low accuracy. Flat starts are only viable when there is a large enough amount of data e.g. over 25 hours of a single speaker in the experiment by Peddinti and Prahallad (2011). The Czech, Hungarian and Russian phone models used in this chapter are non-trivial to adapt. However if they could be easily adapted, it would be interesting to apply these acoustic phonetic refinements to the system. This could potentially combine the advantage of higher accuracy of adaptation, with the small amount of target data needed for cross-language forced alignment.

4.3.1 Phone set similarity

In mapping from one phone set to another in cross-language forced alignment, some languages show a closer mapping to the target language phone set than others. An example of the automatic phone set mapping is shown in Table 4.4 with the distance of each phone mapping indicated by the BFEPP measure.

Sometimes there is more than one phone at a minimal distance to the target phone. For example with the TIMIT phone [ɾ̃] there are actually two close Russian phones [r̪] and [n̪] both with a BFEPP distance of 4 features. The algorithm has been implemented to make a deterministic choice, namely the phone that is closest based on the Unicode code point. In this case it is [r̪]. A less arbitrary way of making the choice would be to use the most salient feature. This could be implemented when there is more data and consensus on this subject, particularly from a language universal perspective.

Phone set	PER	BFEPP
Czech	**56%**	**0.90**
Hungarian	65%	1.16
Russian	69%	1.35

Table 4.5: Expressing the TIMIT phone set: phonetic distance

The closeness of match can be evaluated by how well the target language phone set is expressed with this source language phone set. One way to measure this is to look at the proportion of phones that do not match exactly. This is equivalent to the PER of the mapping (e.g. in Table 4.4 the mapping error is $4/9 = 44\%$ PER). Another way to measure it takes account of how distant each phone is phonetically. This is equivalent to the BFEPP of the mapping (e.g. in Table 4.4 the mapping error is $7/9 = 0.78$ BFEPP). The results of these measures applied to the problem of representing the TIMIT phone set are shown in Table 4.5. This gives an assessment of how well different language phone sets represent the TIMIT phone set.

Both measures suggest that out of the three languages, the Czech phone set is best for representing the US English TIMIT phone set. The ranking of the languages is also the same as the recognition and alignment experiments. This suggests that the closeness of phone sets can be used to predict performance of cross-language phone recognition and cross-language forced alignment. This was attempted in the next experiment.

4.3.2 Cross-language forced alignment on Bosnian (Experiment 4C)

Following publication of the above experiment on TIMIT (Kempton et al., 2011), interest was expressed from a linguist[4] working on conversational analysis of Bosnian. A dataset of 3.8 hours had been recorded from a series of meetings in Bosnian with each speaker in a separate channel. An orthographic transcription had been created that needed to be aligned with the audio. Apart from restricted commercial systems, a speech recogniser for Bosnian is not readily available. The linguist had also made enquiries about training a speech recogniser from scratch to do the forced alignment, i.e. a flat-start. According to an expert in large vocabulary continuous speech recognition, the low quantity and quality of the data (noise and conversational style) made it difficult to predict satisfactory results.[5] Also with limited resources at that time for setting up such a system, it was decided instead to attempt cross-language forced alignment.

The phoneme inventory for Bosnian is well known, so the experiment is slightly different from the previous forced alignment experiment; this time the transcription is phonemic not phonetic. However, it is still useful for showing how the same technique can be used on different languages. The transcription provided for Bosnian was originally orthographic and, because

[4]Kurtic (2011), personal communication
[5]Hain (2011), personal communication

Phone set	PER	BFEPP
Czech	**23%**	**0.55**
Hungarian	50%	0.75
Russian	53%	0.73
TIMIT	37%	0.90

Table 4.6: Expressing the Bosnian phoneset: phonetic distance

Recogniser	20ms Error
Czech	53%
Hungarian	59%
Russian	**51%**
TIMIT	57%

Table 4.7: Cross-language forced alignment on the Bosnian dataset

there is generally a one-to-one correspondence from grapheme to phoneme, the letters were mapped directly to phonemes. Speech was grouped into Turn Constructional Units (TCUs); similar to but not exactly the same as what might more loosely be called utterances. Results are given as the mean performance expected on such an utterance. The same recognisers were used as in the previous experiment, but this time an additional TIMIT recogniser was added which analyses the audio at 16kHz sampling frequency.

It was suggested in Section 4.3.1 that the closeness of phone sets can be used to predict which language will work best for cross-language forced alignment. Therefore it was decided to predict the performance of cross-language forced alignment on Bosnian. This is shown in Table 4.6. Using the BFEPP measure, the Czech recogniser is predicted to perform best and the TIMIT recogniser is predicted to perform worst. Using the PER measure, the Czech recogniser is also predicted to perform best, but the Russian recogniser is predicted to perform worst.

In the Bosnian dataset, there are four people in the conversation; three females and one male. For the evaluation, about 5 minutes of each speaker was manually aligned by an experienced phonetician. Forced alignment was conducted in the same way[6] as described previously in Section 4.3. Results showing the performance of the different recognisers are shown in Table 4.7. The evaluation was primarily to determine the best recogniser for processing the complete 4 x 3.8 hours of meeting data. The resources available for this processing only allowed the use of one recogniser so a combination method of recognisers was not evaluated on Bosnian.

To test for statistical significance, the *related samples Friedman's two-way ANOVA by ranks test* was used because the error values did not show a normal distribution for all the recognisers.

[6]This time SpeechCluster (Uemlianin, 2005) software was used to help with the conversion from Praat TextGrids to HTK label files

The factor *recogniser* had a significant effect ($p < 0.001$). For post-hoc pairwise comparisons the *related samples Wilcoxon signed rank test* was used with the Bonferroni correction. There was a significant difference between the following recognisers: ('>' indicates the first recogniser performing better than the second) Russian > Hungarian ($p < 0.001$), Russian > TIMIT ($p < 0.01$), Czech > Hungarian ($p < 0.01$), Czech > TIMIT ($p < 0.01$).

The results can be compared against the phone set similarity predictions. The BFEPP-based predicted ranking of the four recognisers, appears to be a good prediction. In fact, when only statistically significant differences are taken into account there are no errors in the predicted ranking of all possible pairs. For example, the Czech recogniser was predicted to have the highest performance, followed by the Russian recogniser. The results indicate that it was the other way round with Russian performing best followed by Czech. However there is no statistically significant difference between them; so the prediction is not strictly in error. In contrast the PER-based predicted ranking of the four recognisers makes clear errors. Hungarian is incorrectly predicted to perform better than Russian, and TIMIT is incorrectly predicted to perform better than Czech. It could be argued that the TIMIT recogniser should not be included in this assessment of the predictions because it uses a different bandwidth of audio. However the exclusion of the TIMIT results would still show BFEPP-based prediction to be more accurate than the PER-based prediction.

BFEPP-based predictions for this one experiment is promising but not conclusive. Further experiments are needed to verify whether this prediction based on the different phone sets is consistently accurate in predicting the best language to use. Other predictors of language relatedness might include wordlist comparisons or existing language genealogy knowledge. For example Hungarian is in a completely different language family (Uralic) from both English and Bosnian (Indo-European); and therefore the prediction would be that the Hungarian recogniser is likely to be one of the lower performers on both. However, detailed language genealogy information may not be available in the initial survey of an undocumented language.

Time savings on the Bosnian data

Given the above results in Table 4.7 it was decided that the Russian recogniser would be used for forced alignment on the whole Bosnian data set. This contains four channels of speech each lasting 3.8 hours in total.

In creating the original ground truth files, manual phone alignment took about 3.5 hours of work for 5 minutes of recorded material from each speaker (of which there was about 30 to 50 seconds of actual speech from each speaker) i.e. 40x slower than real time. Automatic alignment, which has not been optimised, was closer to 3x slower than real time. After automatic alignment, the linguist made corrections which took an hour for 5 minutes, i.e. 12x slower than

real time. Here is the feedback from the linguist[7]:

> "I assume that the time would be slightly reduced if I had a training session, as the errors are quite consistent and once you get used to them, the checking of each phone is not necessary any more. So you could go with 40min [for 5 minutes of material] instead of the hour if you like.
>
> The very obvious errors are related to spontaneous speech phenomena, like creaky voice, quick and silent articulations at TCU beginnings and ends, laughter and outbreath overlaid on speech and loud inbreaths, also these seemed to be more frequent in false starts, or short TCUs, not necessarily in longer stretches of 'grammatically' correct talk.
>
> It also seemed that the main misalignments were in shortening the vowels in vowel-nasal/plosive/glide transitions and taking only the duration of the closure to be a plosive; there is usually the release with some aspiration that wasn't included in the plosive. Also things like 'uh' are problematic, but is is generally the question what phones are involved there, as there is no standard orthography.
>
> The longer stretches of talk which constitute 'full sentences' work surprisingly well. I wonder why it's only 49% correct in the formal evaluation, the impression on correcting it is that it would be around 70% at least, but it seems that the above cases are quite frequent in the data."

It was found that the correlation of length and error was negative $\rho = -0.52$ ($p < 0.001$) indicating that longer utterances are aligned more accurately. This confirms the linguists' perception that longer stretches of talk perform better.

Further inspections also showed that the forced alignment algorithm was quite sensitive to pauses. If pauses were indicated in the orthographic transcript e.g. with a comma, performance was much better than when they were not indicated.

Overall using the forced-alignment algorithm and correcting the errors took 30% of the time taken when compared with manual alignment. In fact after correcting the errors in the test sample, it was decided that these errors could be tolerated. So the complete Bosnian dataset went through cross-language forced alignment without correction which took 10% of the time needed when compared to manual alignment. Optimising the scripts and streamlining the process would improve this further. Further information on the Bosnian corpus, and how this technique of cross-language forced alignment was used in the development of this resource, can be found in Kurtic et al. (2012).

[7] Kurtic (2011), personal communication

4.4 Conclusions

In the work reported in this chapter, phone recognisers were evaluated on a cross-language task with minimum target knowledge. Performance was evaluated with the PER (phone error rate) and the arguably better measure, BFEPP (binary feature errors per phone). Since there was a lack of suitable data available in an unwritten language, the TIMIT corpus was used. This was appropriate because with such a large number of allophones, it can be considered that the phoneme inventory is not known. The task of cross-language phone recognition for producing a narrow phonetic transcript is extremely challenging and, similar to a previous study (Walker et al., 2003), there were many errors in the output. Czech, Hungarian, and Russian phone recognisers were used in the experiments and results confirmed the difficulty of cross-language phone recognition: 74% PER and 3.2 BFEPP for the highest performing recognizer. The PER measure is interpreted very strictly, which is one reason why this particular result has a high error rate. Without a detailed analysis of errors, it is difficult to be conclusive in identifying where the limitations are. However there will be many sounds that the recognisers were not trained for and, since context modelling is implicit (Section 4.2.1) , this also applies to novel contexts. For example because of Hungarian vowel harmony constraints, the Hungarian recogniser will come across a number of novel contexts in English that it wasn't trained for. When context is modelled the phone models start becoming similar to phoneme models which, as explained in Chapter 2, don't transfer well between languages. This issue could put a limit on the performance of this approach.

Unlike word-based ROVER it was not possible to improve on the best component recogniser error rate for phone-based ROVER. Instead performance was closer to the average component recogniser error rate. This may mean that there is more stability than using a single component, but further experiments on more target languages would be needed to confirm this. There are two apparent problems with phone-based ROVER. The first is the misalignment of component recogniser outputs, which is affected by the low accuracy of the component recognisers. The second problem is correlated recognition errors, which is affected by the bias towards particular errors in component recognisers. Follow-up tests in Experiment 4A indicated that the first problem was not a major factor behind the poor results. This suggests that a more sophisticated technique for combining recognizer outputs should be used to actively deal with the problem of correlated errors. Also there may be fewer correlated errors as more diverse phone recognisers become available in the future e.g. recognisers based on different architectures.

One of the most promising applications of this study is cross-language forced alignment. This allows any IPA transcript in any language to be aligned with the audio. Cross-language forced alignment enables an under-resourced language to be aligned using a phone recogniser that was trained on a different language. Each phone in the under-resourced language is automatically mapped to the closest phone of the recogniser using the BFEPP distance. It was found that com-

bining the different language phone recognisers by using the median boundary could improve the forced alignment accuracy (Experiment 4B). The accuracy level of the cross-language forced alignment was high enough to be useful to linguists. It allowed a corpus of Bosnian conversation to be automatically aligned using a Russian phone recogniser, reducing by an order of magnitude the time needed when compared to manual alignment (Experiment 4C). Forced-alignment uses the original transcript and this seems to mitigate against the issues of misalignments and correlated errors that affect phone-based ROVER.

For both recognition and alignment, there are some indications that a suitable source language can be chosen based on how close the phone set is to the target language. Further experiments are needed to confirm this.

4.5 Chapter summary

Cross-language recognition often assumes a certain amount of knowledge about the target language. However there are hundreds of languages where not even the phoneme inventory is known. This makes phone recognition more challenging because ideally every possible phonetic contrast that might occur in a language needs to be detected. Following an evaluation of direct cross-language phone recognition, an experiment with the ROVER voting system was attempted. Unlike word-based ROVER it was not possible to improve on the best component recogniser error rate for phone-based ROVER. One of the most promising applications of this study is cross-language forced alignment. This allows any IPA transcript in any language to be aligned with the audio, and was used to create a Bosnian corpus for conversation analysis research (Kurtic et al., 2012).

The significant original contributions of this chapter are as follows:

- Cross-language phone recognition and alignment was evaluated on the TIMIT and Bosnian corpora
- A phone-based rather than word-based ROVER voting system was introduced
- Cross-language forced alignment was introduced with automatic IPA mapping (using BFEPP)
- Combining different language phone recognisers was shown to improve alignment results

Parts of this chapter have been published in Kempton et al. (2011) and Kurtic et al. (2012).

Chapter 5

Complementary distribution

When two different sounds occur in mutually exclusive environments the sounds are described as being in *complementary distribution*. Two sounds that have an allophonic relationship, unless they are in free variation exhibit complementary distribution.

For example, in the Seri language (Marlett, 2005) the sounds [m] and [w̃] both occur. However, the sounds occur in different phonetic environments. [w̃] occurs after [k] when in the same syllable, [m] occurs elsewhere e.g. this is apparent in the word for *woman* [ˈkw̃ɑ̃ːm] and the word for *Seri language* [kw̃ĩkˈiːtom] (Marlett et al., 2005). [m] and [w̃] are in complementary distribution, and alongside other evidence, it can be said that [w̃] and [m] are allophones of the same phoneme . Since [m] has a less constrained distribution than [w̃], [m] is said to be the *elsewhere* or *default* allophone and is the *underlying form*. It is hypothesised that [w̃] is an allophone of /m/.

In this chapter a method for detecting allophones through complementary distribution is evaluated. The method was suggested by Peperkamp et al. (2006), and involves comparing sequential probability distributions with each other using an entropy-based measure. The evaluation was first performed on the TIMIT corpus to simulate an under-resourced language, and then an evaluation was performed on the under-resourced language Kua-nsi. An additional method to identify the default allophone was also evaluated.

5.1 A visual representation for complementary distribution

Existing tools such as Phonology Assistant (SIL, 2008) allow a linguist to search for a range of suspected complementary distributions using a distribution chart. The chart gives counts of particular search criteria in particular environments. However it can be useful to visually represent all the data at once to facilitate the discovery of complementary distributions that may have not been considered. This can be achieved by representing all the phone bigrams

counts in a chart, referred to here as the *phone transition count matrix*. This is simpler but more comprehensive than a distribution chart. The software created to produce this matrix makes use of the SRILM toolkit (Stolcke, 2002) and a spreadsheet to display the data. An example is shown of the Kua-nsi language in Figure 5.1. Since the syllable structure of Kua-nsi is primarily CV not all possible bigrams are shown. [+syllabic] phones are listed along the top, and [–syllabic] phones are listed on the left so that each count is of a particular consonant followed by vowel. For example, the count of five in the top row of numbers indicates the number of times that the syllable [pa] occurs in the word list.

Tones are not included with the phones in order to reduce the complexity of the diagram and it was believed that tone was the least likely feature to interact with the other features, particularly among the consonants. The phones in Figure 5.1 were taken directly from the phonetician's transcription which used particular conventions such as the underline to indicate a retracted tongue root. For example [a̱] could be written in IPA convention as [ạ]. The apical vowel symbol [ɿ] was used which is sometimes written as [z̩] in IPA but [z̩ʲ] or [i̱] is arguably better for describing the position of the tongue body. Also this language has a labiodental articulation for the unrounded vowel [ɪ̱] which could also be written as [ɪ^{ᵛ}]. As in the rest of this thesis, the Phonology Assistant tool (SIL, 2008) was used to assist the sorting of phones in order of articulation, before dividing into [+syllabic] and [–syllabic] groups. The symbols [.] and [#] refer to a syllable boundary and a word boundary respectively.

If there was a completely random sequence of phones, all with the same probability, then the phone transition count matrix would show a uniform distribution. Even with the small sample available from the language (540 words) it can be observed from Figure 5.1 that there are constraints in action. First, it is clear that the vowels [u], [a] and [ɿ] occur frequently. It can be seen that [ɿ] has a very restricted distribution. The retracted tongue root counterpart phone [ɿ̱] exhibits the same distribution. Furthermore it can be seen that for every environment or context where [ɿ] occurs, [i] does not occur, for example after [s]. The same is almost true for [ɿ̱] and [i̱] except for one example where they both follow a glottal stop. The pattern is circled in Figure 5.1 and it is a clear indication of complementary distribution. [i] has a less constrained distribution, is said to be the default allophone and can be hypothesised as an underlying form.

The preceding consonants that appear to trigger the change from the underlying form to the surface form are [t͡s, t͡sʰ, s, d͡z, z]. These sounds belong to the natural class [+strident, –distributed][1]. The change can be written as the following phonological rule:

[1]In the Hayes (2009) feature system, [–distributed] distinguishes these apical sibilants from the laminal-prepalatals e.g. [ɕ]. This also fits with the apical vowel description. Other feature systems may use the [+anterior] feature to make the distinction, which is an important feature in standard Chinese, a language that also exhibits the apical vowel (Duanmu, 2000).

	n̩	ŋ̍	i͡ũ	i͡u	i͡e	i͡õ	i͡a̰	i͡ã	i͡a	i͡ɑu	ɿ̰	ɿ	ḭ	ĩ	i	ɨ̰	ɨ	a͡i	ã̰	a̰	ã	a	y͡a	ɯ̰	ɯ̃	ɯ	ɤ̰	ɤ	u͡o	u͡ɛi	u͡a̰	u͡ã	u͡a	ṵ	ũ	u	õ	o
p									4						2							5				2							2			4		
pʰ				1					4				1		2							9		1		3							2					
p͡f																	9																					
p͡fʰ																	7																					
b				1	1				4				1		4							3				1							4			5		1
b͡v																	3																					
m						1	1		8				1		1	1	11	1		1		3		4	1	3	1	1			1		3		1	19		4
ʔ͡m															1	1				1											1		1					
f																	3											1										
ɱ																	2																					
v																	14																					
t													1		2							5		3		3		2					2			4		3
tʰ				2						1			1		3					1		3				4		2					1			2		
t͡s											2	14								1		2		3		2						1	1			2		1
t͡sʰ											1	33								1		8				4							3			4		2
t͡ɕ					1				1						2					1													2			1		1
t͡ɕʰ													1		3							2	1										1			3		1
s											10	20								2		2		1	1	4						1	7			5		
d									1				3		4		1					6	1	1		2		4	1							2		3
d͡z											2	22										6		1									3			3		1
d͡ʑ															1							5	1													4		1
n													1		3					3		10		2		8		2							1	8	2	2
ʔ͡n													4		1									3		2								4				
z											2	10										2											1			12		
l									1				1		3	1				13		11		1		4		9			1		5	1		6		4
ʔ͡l																				1						1												
ȵ													1		11					1			2											1		3		1
ʔ͡ȵ													3																									
ɕ													1		5								1													5		
ʑ													2		2							1																
ç			1			1			2				1																							1	1	
j															4							3														5		2
ʔ͡j																				2		3												1			1	
k																				2		13				6		1					4			2		4
kʰ																1						6				5	1	4				1	2			4	1	2
x																								1		2				1			4			7		
g																	2			1		6		1		8							1			8		1
ŋ																						2				3						1				3		
ɣ																				1						4							1	1		4		
w																			2	1		5														1		1
ʔ	10	1									1		1		3	1	1				1	30		1											2	13	1	1
h								1						1	4						3	3			3	1						3	1			1	2	
#	3											4					3									1							1			5		
.	3	1							1			1						2				1														1		

Figure 5.1: Phone transition count matrix for Kua-nsi showing consonants followed by vowels. Evidence for complementary distribution between [ɿ] and [i] is circled, showing that the phonetic environments i.e. the preceding consonants are disjoint sets.

$$\begin{bmatrix} -\text{back} \\ +\text{high} \\ -\text{round} \end{bmatrix} \rightarrow \begin{bmatrix} +\text{strident} \\ -\text{distributed} \end{bmatrix} / \begin{bmatrix} +\text{strident} \\ -\text{distributed} \end{bmatrix}\_$$

Other possible complementary distributions are not so clear, but it could be hypothesised that there is also a complementary distribution between [ɿ] and [ɤ]. If this was confirmed to be a second allophonic relationship, then there would be neutralisation occurring[2].

Sometimes correcting or reinterpreting the acoustics will clarify a hypothesis. For example the consonants [x] and [h] are close to complementary distribution but there is a small amount of overlap. When the words containing these sounds were checked again by a phonetically trained listener, the corrections would have confirmed a complementary distribution (e.g. the word for *axe* [hu^{55}.ts^{h}o^{33}] should probably be corrected to [xu^{55}.ts^{h}o^{33}]). Since the phone transition count matrix is automatically derived from the wordlist the linguist can use it in an iterative way. As corrections or reinterpretations are made, true complementary distribution should become more visually prominent in the matrix.

5.2 A visual representation for interpretation

The phone transition count matrix can also be used as a visual representation to assist in the interpretation stage of a phonemic analysis (see Section 1.2). Interpretation occurs after phonetic transcription and before the rest of the phonemic analysis (including the stage of complementary distribution). In a phonetic transcription there can be ambiguous phones, where it is not clear if these particular sounds are consonants or vowels, or whether they are single phones, sequences of phones or merely a transition. In this situation an inventory is built up from the transcription of shortest possible phone forms, and a full phone transition count matrix is produced.

A full phone transition count matrix of Kua-nsi short phone forms is shown in Figure 5.2. Since the sorting order of phones is primarily into consonants (C) and vowels (V), the matrix can be divided into four types of bigrams representing the syllable forms; CC, CV, VC and VV. This is shown in the centre of the figure. The first time this was done with Kua-nsi, the sorting algorithm placed the apical vowel [ɿ] among the consonants. Solely from the pattern of distributions it could be seen that it was out of place and should be among the vowels. Such visual anomalies demonstrate that the phone transition count matrix can help to resolve ambiguities about whether a sound is a vowel or a consonant.

But what about ambiguous sequences? The resulting visualisation indicates a constrained CV syllable structure. There are no codas. There appear to be consonant cluster exceptions but

[2]As reported in Appendix B, it is possible that an alternative transcription for the Kuan-nsi sound [ɤ] is [ɯ] and vice versa. If this is the case and neutralisation is occurring, then the phonological rule stated earlier would be simplified so that the target vowel would not need to be [-back].

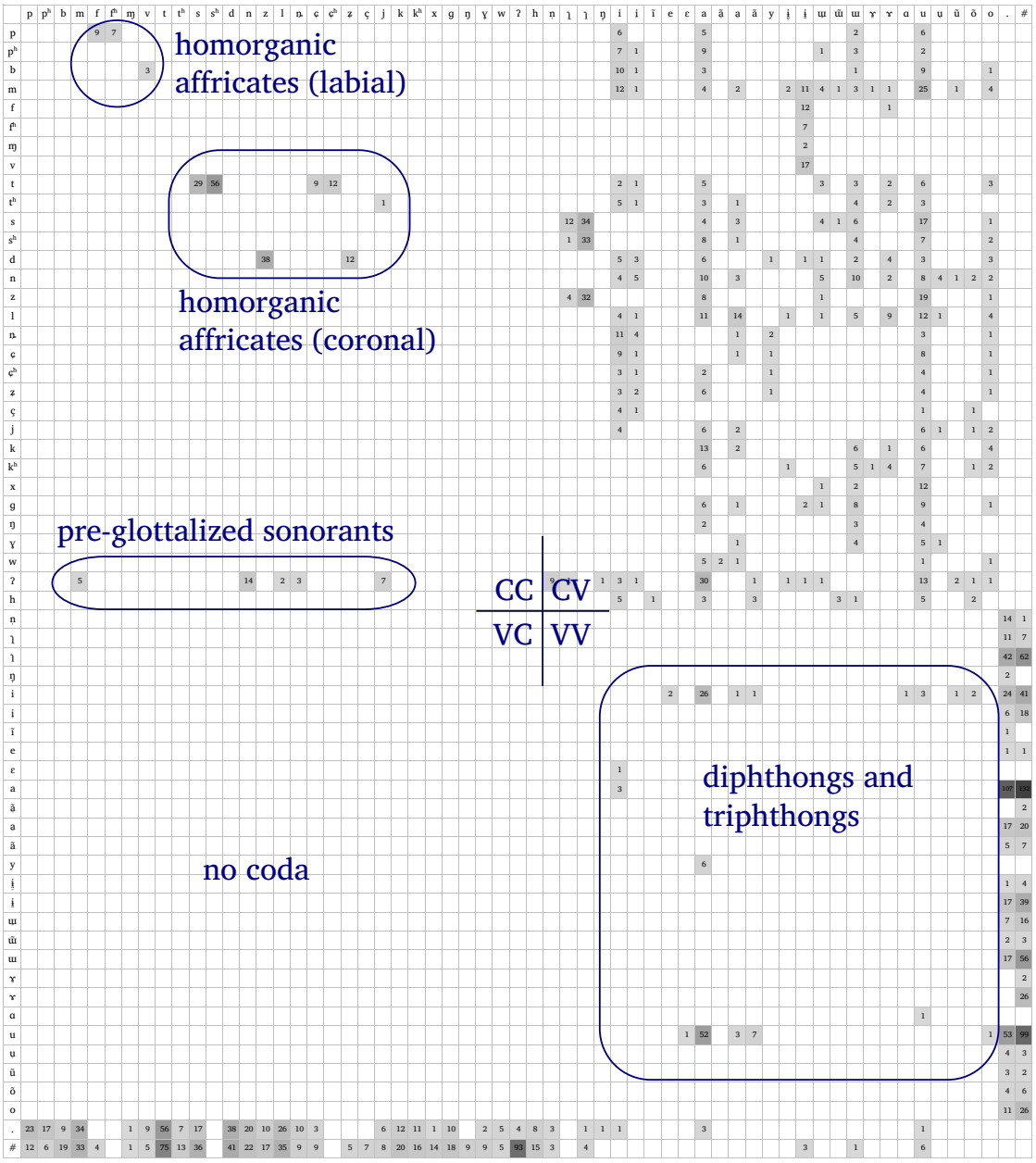

Figure 5.2: Full phone transition count matrix for Kua-nsi with short phone forms for the purpose of assisting in the procedure of interpretation. The cross in the middle separates the grid into consonant and vowel pairs and the hypothesised contour segments are circled.

these are best interpreted as sequences behaving as single phones. The figure helps to show through the way the phones are clustered in the matrix that there is a simple generalisation to these exceptions; they are homorganic affricates and pre-glottalized sonorants. The bottom two rows indicate the start of words and syllables. The right hand side of the bottom two rows show that a few syllables do start with a [+syllabic] phone, indicating a (C)V syllable structure. Since the transcription included syllable boundaries, the vowel sequences can be interpreted as diphthongs or occasionally triphthongs.

This interpretation of Kua-nsi phone sequences was independently confirmed in an analysis by the linguist field workers (Lehonkoski et al., 2010). The wordlist was then automatically rewritten according to this interpretation and then taken as a basis for the rest of the analysis including complementary distribution. This means that zooming in at the top right of Figure 5.2 and applying the interpretation produces the matrix shown in Figure 5.1.

The consonants in the Kua-nsi dataset are used for the evaluation of the different algorithms in the following sections, but as stated in Section 1.4, there was some uncertainty about the vowel system. The patterns discovered from using the phone transition count matrix, were passed back to the field linguists. This influenced the decisions made about the orthography and alphabet[3], e.g. it was decided that the apical vowel [ɳ] would not be represented by a separate symbol because it was the allophone of another underlying form that was already represented in the alphabet.

5.3 Measuring complementary distribution

Peperkamp et al. (2006) proposed the Kullback-Leibler measure of the similarity between two probability distributions to highlight possible complementary distributions. A symmetric version of the measure was used. Kullback and Leibler (1951) originally defined what they call *the mean information for discrimination* as an asymmetric measure commonly now referred to as *relative entropy*. However they also denote a symmetric divergence which they compare with a measure from Jeffreys (1948). This is the sum of both permutations of relative entropy. To avoid any confusion, the symmetric version will be referred to as the *Jeffreys divergence* and the asymmetric version as *relative entropy*. The Jeffreys divergence between the distribution of two phone segments s_1 and s_2 with context c is defined as:

$$D_J(s_1, s_2) = \sum_{c \in C} \left(p(c|s_1) \log \frac{p(c|s_1)}{p(c|s_2)} + p(c|s_2) \log \frac{p(c|s_2)}{p(c|s_1)} \right)$$

From this form of the equation (Peperkamp et al., 2006) it is clear that to calculate the divergence all that is needed is the bigram conditional probabilities. For example to calculate the Jeffreys divergence of [x] and [h] requires the conditional probabilities $p(c|[\text{x}])$ and $p(c|[\text{h}])$.

[3]Crook and Castro (2012), personal communication

A simple way of deriving this is to use the counts directly. For example the probability of the context segment [u] given that segment [x] has just occurred can be estimated as follows:

$$p(c|s) = \frac{n(sc)}{n(s)}$$
$$p([\text{u}]|[\text{x}]) = \frac{n([\text{xu}])}{n([\text{x}])}$$
$$p([\text{u}]|[\text{x}]) = \frac{7}{15}$$

where $n()$ is the count of a particular sequence. The count values are the same as those in Figure 5.1. Since all contexts need to be considered, the counts corresponding to rows [x] and [h] in Figure 5.1 are effectively being compared with each other when calculating Jeffreys divergence of [x] and [h].

From Figure 5.1 it can be seen that the probability estimate $p([\text{u}]|[\text{h}]) = \frac{1}{23}$. As mentioned in Section 5.1 that single occurrence of [hu] comes from [hu^{55}.ts$^{\text{h}}$o^{33}] which should probably be corrected to [xu^{55}.ts$^{\text{h}}$o^{33}]. This would result in a zero probability estimate which would cause problems in the Jeffreys divergence calculation. To deal with this problem, smoothing is used to ensure small but nonzero probability estimates. Peperkamp et al. (2006) used *add-one* smoothing to derive probability estimates. The study here uses the SRILM toolkit (Stolcke, 2002) for estimating transition probabilities from the transcript. The smoothing method is *Katz back-off with Good-Turing discounting* which generally gives better probability estimates than add-one smoothing. Similar to Peperkamp et al. (2006), only the following phone is used as the context for complementary distribution.

5.4 Results on TIMIT (Experiment 5A)

5.4.1 Predicting allophonic relationships

The Jeffreys divergence algorithm applied to the TIMIT data of 1386 utterances is shown in Figure 5.3 with each Jeffreys divergence value rounded to the nearest whole number. Although these values are shown for the consonants, the analysis has also involved taking account of vowels, utterance boundaries and pauses. It can be seen that the highest scoring pair is [j, ŋ] because these phones have quite different environments; [j] for example is most frequently followed by the vowel [ʉ], whereas [ŋ] never appears in this environment but instead is frequently followed by [k] or an utterance boundary. There are other similar examples that show apparent complementary distribution but are not actually allophones.

An alternative but similar measure to the Jeffreys divergence is the Bhattacharyya distance. One of the differences is that the probability estimates do not need to be smoothed. Table 5.1

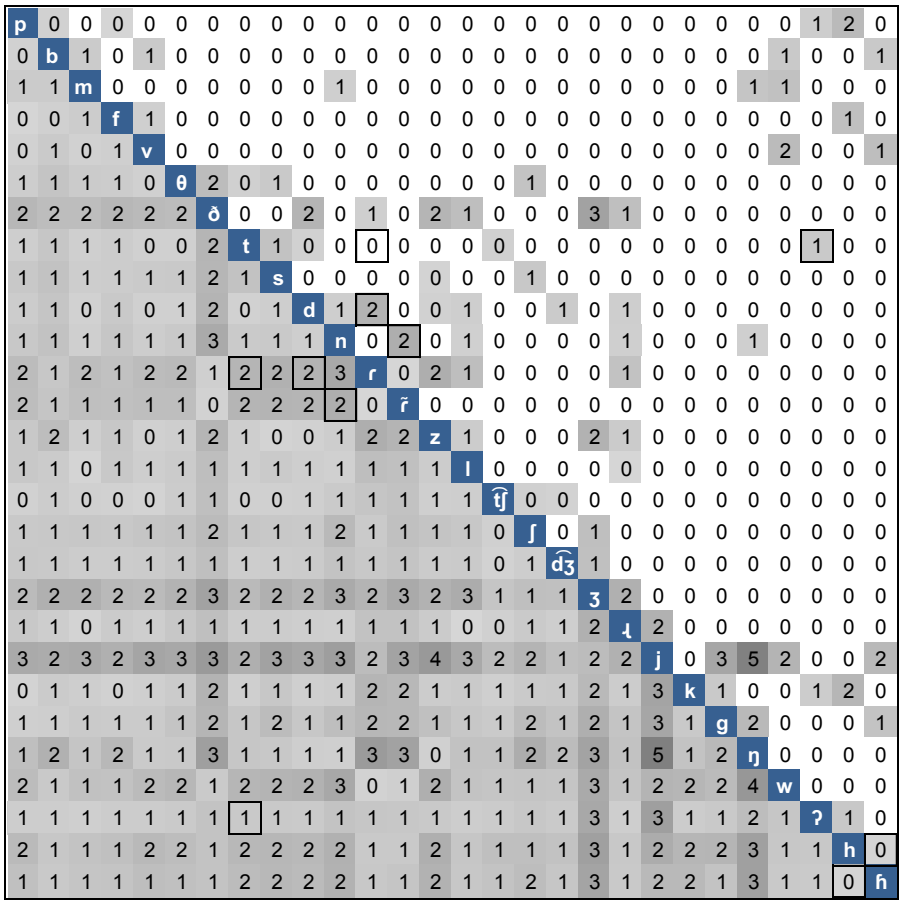

Figure 5.3: Phone relationship chart showing Jeffreys divergence (bottom left) combined with other filters (top right). Each cell represents a phone pair shaded in proportion to the divergence. Values are rounded to the nearest number. Outlines mark actual allophones.

shows that both algorithms had a beneficial effect each with an ROC-AUC result that is greater than chance (50%). In common with Peperkamp et al. (2006) it was found that the Jeffreys divergence performed better than the Bhattacharyya distance.

A comparison of the effectiveness of the Jeffreys divergence algorithm with the phonetic detection algorithms (see Chapter 3) is shown in Table 5.1. A combination of algorithms is also shown, these are combined by multiplication in the same way as Peperkamp et al. (2006) (see end of Section 3.1.2). The ROC-AUC score shows that the Jeffreys divergence algorithm performs quite poorly when compared to the phonetic distance algorithms of Chapter 3. However the Jeffreys divergence algorithm combines well with the other filters, and the results show that each process does make a contribution.

Algorithm applied to TIMIT	ROC-AUC	PR-AUC
Bhattacharyya distance	57.3%	2.0%
Jeffreys divergence (JD)	58.9%	2.3%
JD and active articulator	74.3%	3.6%
JD and relative minimal difference	81.9%	9.4%
JD and both	82.8%	10.7%

Table 5.1: Area under the ROC and PR curves for the different algorithms on TIMIT

5.4.2 Detecting the default phone

Once allophone pairs are found, it can be useful to determine which member of the pair is the default phone. Peperkamp et al. (2006) suggest using relative entropy, where the phone with the lowest relative entropy in association with its context should be regarded as the default phone. This means that the default phone has the least constrained distribution. For a pair of phone segments s_a and s_d the default phone is defined as

$$s_d = \min_{s_a, s_d} \left[\sum_c p(c|s) \log \frac{p(c|s)}{p(c)} \right]$$

The other phone segment s_a is the allophone. This technique identified the correct phone for all five allophone pairs within the TIMIT consonant experiment. This outcome corresponds to a 3% probability of getting this result by chance.

5.5 Comparison with previous studies

The ROC-AUC measure can also be calculated from the results of previous studies. Rather than using ROC curves, this is calculated from single thresholds as described earlier in Section 3.2, using the probabilistic interpretation of ROC-AUC. These ROC-AUC values for previous studies are shown in Table 5.2. The first two rows refer to results from the original study by Peperkamp et al. (2006) on French with the performance of the Jeffreys divergence (JD) algorithm estimated from a threshold (set as one standard deviation above the mean value). Both the relative minimal difference and an assimilation criterion filter are used (see Section 5.7.1). The other results refer to the study of Le Calvez et al. (2007) on French and Japanese. There are two important differences in this study; first the Jeffreys divergence algorithm took into account both contexts of the following phone and the preceding phone. Second, a reliability filter is included to discard pairs that are not statistically reliable. To avoid confusion it should be noted that these previous studies referred to the Jeffreys divergence as the Kullback-Leibler measure.

Algorithm and dataset	Recall	False alarm rate	ROC-AUC
French-2006: JD	0.778	0.138	82.0%
French-2006: JD & RMD & AC	0.778	0	88.9%
French-2007: JD2 & R	0.727	0.433	64.7%
French-2007: JD2 & R & RMD & AC	0.727	0.001	86.3%
Japanese-2007: JD2 & R	0.533	0.522	50.4%
Japanese-2007: JD2 & R & RMD & AC	0.533	0.001	76.6%

Table 5.2: Area under ROC curve for previous studies; JD = Jeffreys divergence, JD2 = JD with both contexts, RMD = Relative minimal difference filter, AC = Assimilation criterion filter, R = Reliability filter.

Peperkamp et al. (2006) showed a better performance for the JD algorithm than in this current study on TIMIT. This may be because their corpus was much larger (42,000 utterances versus 1386 utterances in TIMIT). However, evidence earlier in Peperkamp et al. (2006) where corpus size is studied, suggests that a corpus the size of TIMIT is not too small for the JD algorithm to work effectively. The lower score on TIMIT could be indicative of a more challenging corpus in general. This is backed up by the results on French in Experiment 3E. In Le Calvez et al. (2007) the JD algorithm (with reliability filter) takes both contexts into account but this scores lower. Since the ROC-AUC figures are only for the reported threshold levels, they may be slightly higher for the ranked JD values especially if the threshold point is not optimal for the specific experiment. In all the experiments, the addition of the phonetic filters improves results dramatically. Interestingly the ROC-AUC figure for Japanese (50.4%) reveals that, without any of the additional phonetic filters, the JD algorithm with the reliability filters is barely performing better than chance on the language. This is confirmed by calculating the precision and corresponding chance value from the original paper (Le Calvez et al., 2007) giving two very similar values of 1.12%, and 1.09% respectively. Since the reliability filter did not make any difference to the French result (Le Calvez et al., 2007), it appears that it has not been demonstrated to have any beneficial effect on real languages.

5.6 Experiments on Kua-nsi (Experiment 5B)

The experiments performed on TIMIT for complementary distribution were performed on the Kua-nsi data. The results are shown in Table 5.3. It can be seen that the accuracy of the algorithms are ranked in the same order as in the TIMIT experiments. The results on Kua-nsi show a higher ROC-AUC value than TIMIT. The lower PR-AUC reveals a greater number of false alarms which is to be expected because of the greater number of phones in the corpus.

Algorithm applied to Kua-nsi	ROC-AUC	PR-AUC
Jeffreys divergence (JD)	61.8%	1.1%
JD and active articulator	83.6%	3.0%
JD and relative minimal difference	85.6%	5.5%
JD and both	87.7%	8.4%

Table 5.3: Area under the ROC and PR curves for the different algorithms on Kua-nsi

Relative entropy was used to predict which phone in each phone pair was the default phone, as in Section 5.4.2. In the Kua-nsi data there was some uncertainty in the ground truth. The current human-produced phonemic analysis of Kua-nsi is not yet fully mature, and it is not yet known which phone in the pairs [h,x] and [ʑ,j] is the default phone. The relative entropy algorithm predicted that [h] and [ʑ] were the default phones respectively. For the four phones that were certain, all were correctly identified.

5.7 Feature-based algorithms

The Jeffreys divergence algorithm treats all phones as arbitrary symbols and has no knowledge of their features. And yet, as seen in Section 5.1, features are especially relevant to the sequential constraints imposed on groups of phones in a particular language. Since phonology rules commonly apply to natural classes (Hayes, 2009, p.71), it is important to integrate features into the algorithms for detecting allophones, and this is what is investigated in this section.

5.7.1 Assimilation (Experiment 5C)

An assimilation detector was introduced in Peperkamp et al. (2006) where it was referred to as a filter for allophones. As with the other filters used by Peperkamp, only combined results with the other algorithms were given. In this chapter standalone results are given. Peperkamp defines an assimilation criterion based on the premise that an allophone should be phonetically closer to its context than the default (elsewhere) phone i.e. it should show more assimilation. A possible allophone is confirmed by testing whether for every single feature the total difference summed over the allophone's contexts is less than or equal to the total difference with the default phone. In the original definition of this detector, *context* refers to the following phone. This detector does not work well on the TIMIT data using the Hayes feature set (for information on this feature system see Section 3.1.1) because there is an incompatibility with the feature set used. In the Hayes features a tap is given its own natural class i.e. it has the feature [+tap]. The allophone [ɾ] of /d/ is therefore usually recognised as more distant to its contexts than would normally be assumed to be the case. This is the reason for the poor result on the first

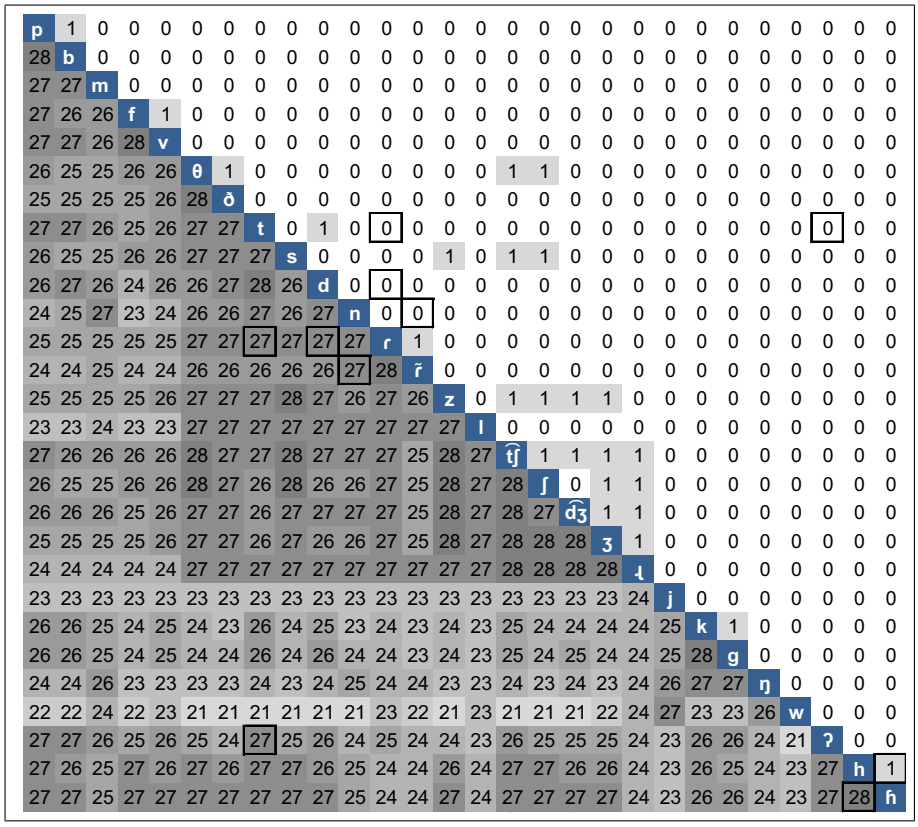

Figure 5.4: Phone relationship chart for Experiment 5C on TIMIT showing the assimilation criterion (top right) and the assimilating features algorithm, where the values indicate the number of features that satisfy the criterion (bottom left). Outlines mark actual allophones.

line of Table 5.4 which shows the results for Experiment 5C. The specific values for phone pairs is shown on the top right triangle of Figure 5.4. There is also a more general limitation with this detector, as the authors state (Peperkamp et al., 2006); it is not completely universal. For example in RP English the clear and dark L allophone pair [l, lˠ], do not show strong assimilation with their environments, particularly in regard to the position of the tongue body (cf. Sproat and Fujimura, 1993).

The assimilation algorithm, however, can still be used to assign a certain confidence level to allophone pairs rather than making a hard decision. The original requirement that every single feature must satisfy the assimilation criterion can be relaxed. Instead the number of features that satisfy the criterion is given as a score. This is shown in the bottom left triangle of Figure 5.4 and the second line of Table 5.4. This change to the algorithm allows it to be more robust to different feature conventions e.g. as described above. This certainly makes a difference with the performance on TIMIT; with the tap feature being handled more appropriately. There is a

Algorithm and dataset	ROC-AUC	PR-AUC
TIMIT Assimilation criterion	56.7%	2.3%
TIMIT Assimilating features	83.9%	4.0%
Kua-nsi Assimilation criterion	74.4%	2.2%
Kua-nsi Assimilating features	77.1%	2.3%

Table 5.4: Area under the ROC and PR curves for the different feature-based assimilation algorithms

positive but smaller effect on the results of the Kua-nsi data.

Overall it can be seen that a knowledge of features is beneficial.

5.7.2 Towards a probabilistic feature based framework

Given the limited amount of data that is available for a phonological analysis, it is not always easy to notice when a particular distribution is constrained. For example, in Figure 5.1, without knowledge of other distributions or features the distribution of [i] following other phones could easily be interpreted as uniform. However, once the natural class [+strident,–distributed] is taken into account, a clearer constraint emerges that [i] never follows these sounds.

This more obvious constraint can be measured by the entropy of the probability distribution. For example, consider the phones preceding [i]. This is given by the probability distribution f_X:

$$f_X = p(phone_{t-1}|[\text{i}]_t)$$

The values for these probabilities, as shown in Figure 5.1, can be used to calculate the entropy:

$$H(X) = 4.04 \text{ bits}$$

Once the natural classes are taken into account there is a very simple probability distribution f_Y:

$$p(\begin{bmatrix} +\text{strident} \\ -\text{distributed} \end{bmatrix}_{t-1} |[\text{i}]_t) = 0$$

$$p(\neg\begin{bmatrix} +\text{strident} \\ -\text{distributed} \end{bmatrix}_{t-1} |[\text{i}]_t) = 1$$

$$H(Y) = 0 \text{ bits}$$

A low entropy means a strong constraint on the distribution. Calculating the entropy for different groupings of features can be used to detect differences between distributions; and to indicate which features are involved in the phonological rule.

A promising framework for modelling the probabilities of features are factored language models (Kirchhoff et al., 2008) which allow a number of conditional probabilities to be modelled between features and phones. This allows a comparison of two phones and testing for the influence of particular features. It could also be used in a more unsupervised fashion, by automatically finding the most optimum dependencies in the language between features and phones. In this way a large part of a phonology of the language could be modelled to produce a language model with low entropy (and therefore a low perplexity). However this unsupervised approach is nontrivial, with Kirchhoff et al. (2008) recommending genetic algorithms to optimise a language model for such a large search space.

5.8 Conclusions

The phone transition count matrix is very helpful for visualising the phonology of a language in a single image. This is useful both for interpretation, and for discovering complementary distributions. The ordering of the phones is an important part of visualisation. The order used here is essentially the same order used in Phonology Assistant (SIL, 2008) which is sorted by articulation (from front to back) and also split into consonants and vowels. A few modifications were made to the latter grouping to ensure [+syllabic] and [–syllabic] categories with syllabic consonants occurring in between. The resulting visualisation at the macro level immediately indicates the syllable structure and at the micro level there is an indication of complementary distribution and phonotactic constraints. It is believed to be the first time that this type of visualisation has been used to assist linguists.

The results in this chapter show that the application of the Jeffreys divergence algorithm introduced by Peperkamp et al. (2006) can help detect allophones among the consonants in the TIMIT (Experiment 5A) and Kua-nsi (Experiment 5B) corpus. These are challenging corpora where the transcriptions are more faithful to the acoustic signal than in past experiments. It is not surprising, therefore, that some performance figures are lower than in previous studies that were conducted in more ideal conditions (Peperkamp et al., 2006; Le Calvez et al., 2007) (see also Experiment 3E in Section 3.4.4). Using the ROC-AUC measure, the Jeffreys divergence algorithm was shown to be less effective than the phonetic similarity algorithms. In common with previous studies, results for the JD algorithm are also reported for combined experiments with phonetic distance detectors. The JD algorithm made a contribution when combined with other algorithms, but the phonetic distance detectors made the biggest contribution.

As in previous work (Peperkamp et al., 2006), it was found there were many apparent complementary distributions that were not allophones. This appears to be the main reason the

algorithm performs poorly. Complementary distributions that are not related to allophones, are often due to constraints associated with syllable structure. One extreme example of this, in many languages, is of vowels that are in complementary distribution with consonants.

Once the allophone pair had been detected, the relative entropy measure was able to identify the default phone correctly. Although there was limited data, this did appear to work well. This was a more constrained problem than complementary distribution, since it was already known that the phones had an allophonic relationship.

The work here had a similar focus of scope to Peperkamp et al. (2006) because the investigation was on the distribution of the succeeding environment rather than the preceding environment. This could be easily extended to a similar investigation of the preceding environment, and potentially to both environments although a previous study has not shown that this is particularly beneficial to date (Le Calvez et al., 2007). This could be a modelling issue, where the search space becomes too sparse for effective generalisations. However the better results in modelling the succeeding environment could be evidence of the dominance of anticipatory processes in articulation.

The feature-based assimilation algorithm adapted from Peperkamp et al. (2006) gives much better results than Jeffreys divergence algorithm (which doesn't take into account of features). The ROC-AUC value of 83.9% is the highest achieved on TIMIT although the PR-AUC value and the results on Kua-nsi show that it is not consistently superior than the BFEPP phonetic distance measure. Feature-based algorithms seem to be the most promising direction for detecting the type of constraints that are manifested in complementary distribution. This demonstrates the significance of features in allophony, and further experiments with a feature-based model may help to reveal a better model for modelling the phonetic/phonological phenomenon underlying complementary distribution.

5.9 Chapter summary

In this chapter a method for detecting allophones through complementary distribution is evaluated. This was principally the Jeffreys Divergence algorithm, adapted from Peperkamp et al. (2006); it did not make use of features, and performed relatively poorly. Further algorithms related to complementary distribution were also evaluated. The assimilation criterion also adapted from Peperkamp et al. showed poor a performance on TIMIT; this appeared to be due to an incompatibility with the feature set used. The diagnosis of this problem led to the development of the assimilating features algorithm that performed better on both corpora. Once allophone pairs had been discovered the relative entropy algorithm correctly identified the default phone. The phonology visualisations developed demonstrate their effectiveness for discovering many phonological patterns. There is much potential for algorithms to take advantage of identifying these patterns, particularly in using features to discover complementary distribution.

The significant original contributions of this chapter are as follows:

- Three complementary distribution related algorithms were evaluated on the TIMIT and Kua-nsi corpora
- The phone transition count matrix was introduced to visualise the phonology of a language
- Peperkamp's assimilation algorithm was adapted for Hayes' feature set and improved

Parts of this chapter have been published in Kempton and Moore (2009).

Chapter 6

Minimal pairs

The use of minimal pairs is regarded as a particularly effective method in phonemic analysis and the only method to conclusively establish contrast between sounds (Hayes, 2009, p.34). In this chapter minimal pairs are quantitatively evaluated for their effectiveness in a phonemic analysis.

There is some variation in the literature regarding the definition of a minimal pair. The following definitions use three different terms; sound, segment, phoneme:

> "[pair] of words which differ by only one sound"
>
> (Ladefoged and Johnson, 2010, p.35)

> "two different words that differ in exactly one sound in the same location"
>
> (Hayes, 2009, p.34)

> "pair of words whose members differ by one segment only"
>
> (Gussenhoven and Jacobs, 2005, p.108)

> "pair of distinct words differing solely in the choice of a single segment"
>
> (Odden, 2005, p.335)

> "pair of words that are identical except for one phoneme, occurring in the same place in the string"
>
> (Fromkin and Rodman, 1998, p.530)

> "pair of words differing in only one phoneme"
>
> (Clark et al., 2007, p.92)

These definitions are arranged in order, so that the broadest definition is at the top. It is considered that *sound* is the broadest term, and that *phoneme* is the most specific term and also

the only term to suggest a contrastive unit. The word *segment* can be regarded as synonymous with the word *phone* as defined in Section 1.5.

It seems clear from the literature that the principle of *same location* within a word is implied by the definitions that don't mention it. But what about the broad definitions that use the term *sound*? Do they imply a narrow definition such as a segment or phoneme? Hayes, for example, often uses the term *phoneme*, but not in this definition. This is because he also refers to minimal pairs that differ by a suprasegmental features such as tone (Hayes, 2009, p.291) i.e. the word *sound* includes segments and suprasegmentals.

The remaining definitions refer to a unit of an unspecified contrastive nature: the *segment*, and the contrastive unit: the *phoneme*.

In a phonemic analysis, where two segments need to be compared, it is not initially known whether they are phonemes or not. But as soon as a genuine minimal pair is found, contrast is established, and the difference between the two words is one phoneme. This process however assumes that there have been no errors or uncertainties in deriving the segments in the first place. In real conditions, particular in survey collections the data is noisier. With noisy data it is perhaps better to refer to *putative* minimal pairs and view these pairs as evidence for contrast rather than being used as the gold standard.

In this Chapter the definition of minimal pair from Odden (2005, p.335) is the most relevant. This is because the phonological framework is segmental (Section 1.4), and minimal pairs are sought for before phonemes have been established. In the experiments phonetic transcriptions are used that cannot be guaranteed to be free of errors, so the expression *putative minimal pair* will be used rather than minimal pair.

The structure of this chapter is as follows: The Kua-nsi dataset is used for the first part of the evaluation. The existence of putative minimal pairs in Section 6.1 is used to predict whether phones contrast, or are likely to be allophones. Then Sections 6.2 and 6.3 looks at putative minimal pair counts. These experiments are then performed on the TIMIT dataset in Section 6.4.

6.1 Existence of putative minimal pairs (Experiment 6A)

To find the putative minimal pairs in a word list, the software Minpair (Poser, 2008) was used. As input, the software takes a wordlist with a column containing a phonetic transcription and a column containing a word identifier, e.g. a translation into English.

6.1.1 Putative minimal pairs in Kua-nsi

For Kua-nsi the wordlist is from Castro et al. (2010). The output from the software is a list of all possible minimal pairs.

It is possible to specify contour segments (see Section 3.3) for the minimal pairs, and this can also be used to approximately model suprasegmentals e.g. tone in Kua-nsi. The following output gives the minimal pairs contrasting the low falling tone and the mid tone modelled by the contour segments [a^{21}] and [a^{33}]. Every possible minimal pair is listed, which is why words are repeated in the list.

[za^{21}]	to descend	[za^{33}]	to hit (a target)
[wa^{21}]	to grow (up)	[wa^{33}]	to write
[wa^{21}]	big	[wa^{33}]	to write
[na^{21}]	early	[na^{33}]	to look
[na^{21}]	wolf	[na^{33}]	to look
[na^{21}]	early	[na^{33}]	to cure
[na^{21}]	wolf	[na^{33}]	to cure
[na^{21}]	early	[na^{33}]	black
[na^{21}]	wolf	[na^{33}]	black

The complete output from the software includes all phone pairs for which putative minimal pairs can be found. Given the standard definitions for minimal pairs and phonemes, normally the existence of a single minimal pair would be seen as establishing contrast between the two relevant phones. In the experiments reported in this chapter the output is regarded as a list of putative minimal pairs. Where a putative minimal pair is found it is considered to be a strong indication that the relevant phones contrast.

The results for the existence of putative minimal pairs among the consonants in Kua-nsi are shown in the top row of Table 6.1 and an explanation for how the evaluation measures apply are given below. The performance is better than chance but relatively low when compared to algorithms in previous chapters.

6.1.2 Evaluating the detection of phonemically distinct phones

The evaluation measures ROC-AUC and PR-AUC described in Section 3.2.1 are for evaluating how well an algorithm identifies allophones. However given the probabilistic interpretation of the ROC-AUC measure; it can be seen that the exact same measure can be given to quantify how well the algorithm identifies phonemically-distinct sounds. Recall that ROC-AUC can be interpreted as the probability that a randomly chosen target (allophone pair) will have a higher score than a randomly chosen non-target (non-allophone pair). Since a non-allophone pair is a pair that is phonemically-distinct and an allophone pair is one that is not phonemically-distinct, the sentence can be reworded. ROC-AUC can be interpreted as the probability that a randomly chosen non-target (phonemically-distinct pair) will have a lower score than a randomly chosen target (non-phonemically-distinct pair).

Algorithm applied to Kua-nsi	ROC-AUC	PR-AUC
Putative minimal pair (MP)	64.3%	1.1%
MP counts	64.0%	1.1%
MP independent counts	66.0%	1.1%

Table 6.1: Performance of minimal pair algorithms on Kua-nsi

Another way of understanding this is that ROC-AUC measures how well a list of targets and non-targets are sorted. So the value quantifies both the success of targets (allophone pairs) given high scores and non-targets (phonemically-distinct pairs) given low scores. This is how the ROC-AUC measure can be interpreted in all the experiments in this thesis. The PR-AUC measure however is different and should not be interpreted in the same symmetrical way.

It is said that a lack of minimal pairs does not prove much (Hayes, 2009, p.35), however a pair of sounds that show no evidence of contrast are more likely to be allophones than a pair of sounds that do show evidence of contrast. For some pairs of sounds i.e. common sounds rather than rare sounds, the lack of a minimal pair can be a salient indication of an allophonic relationship.

Given the above reasoning, the ROC-AUC and PR-AUC evaluation measures are used in this chapter in the same way as previous chapters.

6.2 Counts of putative minimal pairs (Experiment 6B)

A search of minimal pairs on the Kua-nsi corpus comparing the sound [u^{55}] with a nasalised version [ũ^{55}] produced the following output:

[ʔũ^{55}.n̪u^{33}] breast [ʔu^{55}.n̪u^{33}] milk

The putative minimal pair above is actually likely to be the same word. Apart from the semantic relatedness, there are two further reasons. First, in closely related dialects the words are the same (Castro et al., 2010, p.69) second, there is no other evidence of nasalisation being contrastive for vowels in this dialect. Further, when these words were checked again by a phonetically trained listener, it was recognized that the word for *milk* did also have a nasalised vowel in initial syllable; confirming that it was the same word.

With small errors in the transcript being a real possibility, the number of minimal pairs found can provide further confidence that the contrast is genuine, because there is less chance of multiple transcription errors occurring in multiple minimal pairs.

For implementing this simple minimal pair counts algorithm, each phone pair's score is the count of putative minimal pairs found, multiplied by -1. This was done to remain consistent

with the convention that a low score suggests a contrast between a phone pair. The results for Kua-nsi consonants are shown in the second row of Table 6.1. Surprisingly there was no improvement on the previous result, when the number of putative minimal pairs was not taken into account.

On investigating this poor result, it was found that while a number of contrasting sounds had a single putative minimal pair; two sounds that were thought to have an allophonic relationship [x,h] were showing two putative minimal pairs:

[hu͡a^{33}]	thirsty	[xu͡a^{33}]	dry
[hu͡a^{33}]	thirsty	[xu͡a^{33}]	to tear

Similar to the earlier example, on re-listening to these words it was recognised that the word [hu͡a^{33}] should have been transcribed as [xu͡a^{33}], i.e. Kua-nsi for *thirsty* and *dry* are one and the same word.

6.3 Using independent counts (Experiment 6C)

Because of the way minimal pairs are counted, a single transcription error can lead to multiple putative minimal pairs. It is better to count the minimal pairs so that each one is based on separate words i.e. independent transcriptions. If this was the case, the above example would only count as one putative minimal pair, and the first example showing a list of minimal pairs to distinguish tone would count as five putative minimal pairs instead of nine.

This method of counting independent words, was implemented as a post-process to the Minpair software. The results for Kua-nsi consonants are shown in the bottom row of Table 6.1. The ROC-AUC measure shows an improvement over those previous results. The secondary PR-AUC evaluation measure doesn't contradict this but shows smaller changes (see Appendix A).

6.4 Experiments on TIMIT (Experiment 6D)

Following the minimal pair experiments on Kua-nsi, the same algorithms were evaluated on the TIMIT corpus. A wordlist, that is a narrow phonetic transcription of each word alongside an orthographic label, was extracted from the TIMIT corpus. As in Chapter 5, the 1386 phonetically diverse sentences from the training subset of TIMIT were used in this experiment. Phonetic transcripts were converted to IPA. The wordlist was then created by matching up the time-aligned word transcripts with the time-aligned phone transcripts. As is clear in Figure 6.1 these are not isolated words and, although this is not inconsistent with the early definitions of minimal pairs, there will be the impact of connected speech processes which is discussed below. For words with multiple pronunciations the most common pronunciation was chosen. For example,

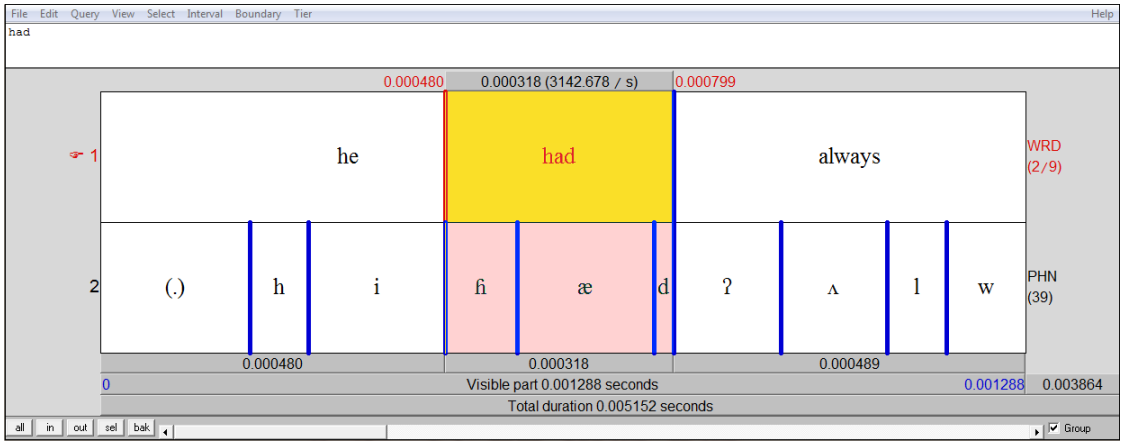

Figure 6.1: Extracting pronunciations from TIMIT was achieved simply by matching the word transcripts with the phone transcripts, e.g. the word “had” in sentence SI2251.

Algorithm applied to TIMIT	ROC-AUC	PR-AUC
Putative minimal pair (MP)	47.5%	1.2%
MP counts	55.9%	1.4%
MP independent counts	53.7%	1.3%

Table 6.2: Performance of minimal pair algorithms on TIMIT (the most common pronunciation is used for each word)

there were 42 instances of the word *had*, 16 different pronunciations, and the most common pronunciation [ɦɛd] was used in the experiments. The resulting wordlist contained 4078 unique words.

The full set of results for TIMIT consonants are shown in Table 6.2. It might be expected that with many more words present, the minimal pair method would show more success on the TIMIT dataset than the Kua-nsi dataset. Surprisingly however the results show that the minimal pair algorithms were, in general, performing little better than chance.

The phone relationship chart shown in Figure 6.2, can help to diagnose the problem. The bottom left part of the chart shows the count of putative minimal pairs which corresponds to the second row of Table 6.2. The top right part of the chart shows the count of putative minimal pairs based on independent words which corresponds to the third row in Table 6.2. For the chart, the shading is consistent with the phone relationship chart in previous chapters; a darker shade if the pair is predicted to be in an allophonic relationship, a lighter shade if the pair is predicted to be phonemically distinct.

It can be seen that the less common phones, such as [ʒ] and [ɾ], rarely form minimal pairs

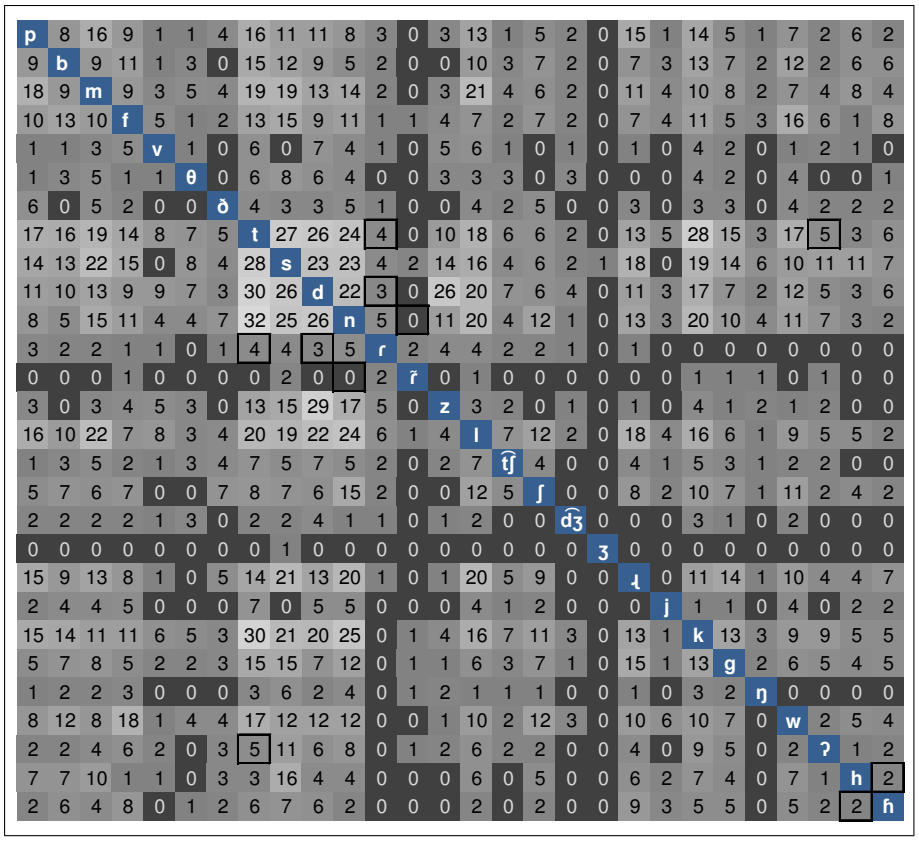

Figure 6.2: Phone relationship chart showing number of putative minimal pairs counts (bottom left) and independent counts (top right). Outlines mark actual allophones. The lighter the shading, the more likely that the phones are phonemically distinct.

Phone pair	Word 1		Word 2	
[t, ɾ]	[gɻe͡ɪt]	great	[gɻe͡ɪɾ]	grade
[t, ɾ]	[ɻa͡ɪt]	right	[ɻa͡ɪɾ]	ride
[t, ɾ]	[sa͡ɪt]	site	[sa͡ɪɾ]	side
[t, ɾ]	[mit]	meat	[miɾ]	meet
[d, ɾ]	[ɻɛd]	red	[ɻɛɾ]	spread
[d, ɾ]	[ɦɘd]	heard	[ɦɘɾ]	herd
[d, ɾ]	[sɛd]	said	[sɛɾ]	set
[t, ʔ]	[kæt]	cat	[kæʔ]	can't
[t, ʔ]	[tɛn]	ten	[ʔɛn]	end
[t, ʔ]	[te͡ɪm]	tame	[ʔe͡ɪm]	aim
[t, ʔ]	[te͡ɪbl̩]	table	[ʔe͡ɪbl̩]	able
[t, ʔ]	[to͡ʊ]	toe	[ʔo͡ʊ]	oh
[h, ɦ]	[hɛd]	head	[ɦɛd]	had
[h, ɦ]	[ho͡ʊl]	whole	[ɦo͡ʊl]	hole

Table 6.3: Problematic putative minimal pairs in TIMIT that appear to be showing contrast between phones that should be allophones according to the TIMIT documentation

with the other phones. This is not surprising, but given that the algorithm does not attempt to deal with this, they are given a score equal to more definite allophones; which could contribute to the false alarm rate. The major problem however is the number of known allophones that have putative minimal pairs. In fact it is only the allophones [n, ɾ̃] that do not have any corresponding putative minimal pairs. The putative minimal pairs for sounds described as allophones in the TIMIT documentation (Garofolo et al., 1993) are listed in Table 6.3.

The false putative minimal pairs arise for a number of reasons. The minimal pairs between [t, ɾ] and [d, ɾ] which involve neutralisation, appeared to be primarily caused by connected speech processes. The difference is consistently in the word final position, and on investigation, the flap was frequently followed by a word initial vowel in the next word. The unusually reduced form for the word *spread* was actually caused by a rare alignment error in the TIMIT corpus. The minimal pairs for the phones [t, ʔ], consistently differ in the word initial position and this is largely down to an interpretation issue; each glottal stop vowel sequence might have been interpreted more appropriately as a single pre-glottalized vowel phone. The minimal pairs for [h, ɦ] only have a difference in the world initial position, but there appears to be no obvious contextual effect from the previous word. In agreement with the TIMIT documentation it was observed that [ɦ] was "typically found intervocalically" (Garofolo et al., 1993) however for the two minimal pairs above there was no such pattern e.g. the voiced glottal fricative appearing after a voiceless stop; *"what had been"* [wʌt ɦɛd bɪn]. Regarding the final example in Table 6.3,

there is also some consistency in the realization of some morphologically related words; *holes* [ɦo͡ʊlz] and *wholesome* [ho͡ʊlsəm]. This suggests some genuine underlying difference, but it is difficult to be conclusive.

Hayes (2009, p.35) explains that "two sounds that appear in a minimal pair are almost always distinct phonemes", and gives two exceptions under the category of *pseudo-minimal pairs*. One exception occurs when distinctions are caused by differences in phonological boundary locations such as word boundaries (Hayes, 2009, p.207). The other exception occurs with *displaced contrasts*, where there is a certain distinction in the underlying form manifested differently in the surface minimal pair (Hayes, 2009, p.146) e.g. a contrast in vowel duration or quality being affected by an underlying difference in consonant voicing.

Clearly putative minimal pairs that turn out not to be minimal pairs are not just due to errors in the transcription. As well as the causes mentioned above, the effect could also be caused by free variation, dialect/idiolect differences, speech rate, and word frequency effects. In the initial stage of a phonemic analysis, it is not known whether a minimal pair is genuine or whether it is a pseudo-minimal pair. So the expression *putative minimal pair* does appear to be a helpful broad term to refer to any minimal pair derived from the narrow phonetic transcript.

6.5 Future work: semantic analysis and rare phones

In Section 6.2, there were two examples in Kua-nsi where a putative minimal pair turned out to be slightly different transcriptions of the same word. The Chinese used to elicit the original data used different words but it appears there is only one word for each pair in Kua-nsi. One example was the word pair *breast* & *milk*, another example was *thirsty* & *dry*.

These anomalies could theoretically be detected by measuring the semantic relatedness between words. One such measure is the explicit semantic analysis algorithm (Gabrilovich and Markovitch, 2007). As a proof of concept an implementation of this algorithm described in Zesch and Gurevych (2010) was used to calculate the semantic relatedness for every possible word pair in the Kua-nsi corpus. The processing was performed by the first author of the above paper.

Part of the output is included in Figure 6.3 which shows the semantic relatedness score for each word pair in the Swadesh 100 list (Swadesh, 1971), a subset of the Kua-nsi wordlist. The first ten words were not included in the analysis, because very common words are currently not included in the explicit semantic analysis index[1]. Results show that while some similar words (e.g. *tooth* & *bone*) show a high semantic relatedness score, other words that might be regarded as opposites and are unlikely to be the same word in any language were also given a high score (e.g. *black* & *white*, were given a higher semantic relatedness score than *black* & *night*).

[1] Zesch (2011), personal communication

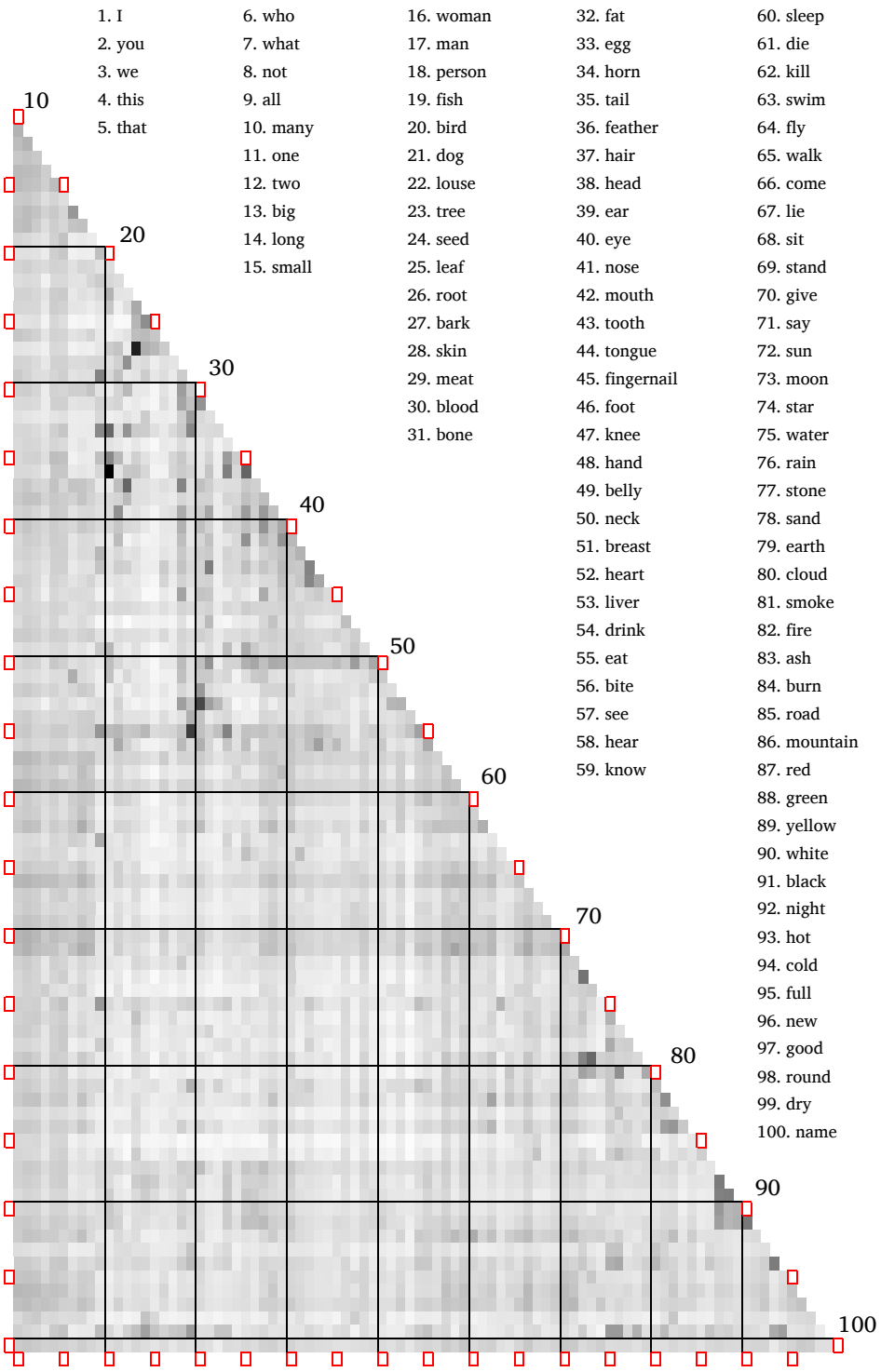

Figure 6.3: Word relationship chart showing explicit semantic analysis scores for the Swadesh 100 list. Darker shading indicates that the word pairs are more semantically related than lighter shaded pairs. The first ten words were not included in the analysis.

It is possible that semantic relatedness for these different words or concepts could vary depending on the language and culture. There will be limits to this and initial experiments by Hassan and Mihalcea (2009) have shown promising results on cross language semantic relatedness.

For future work, it might also be helpful to exclude zero minimal pair counts that involve rare phones. For example in the TIMIT dataset (see Figure 6.2) the phone [n] is very frequent and forms many minimal pairs so it is of significant interest that it forms no minimal pairs with [ɾ̃]. The phone [ʒ], however is rare and it is of less interest that there are many sounds that it does not form a minimal pair with.

6.6 Conclusions

The ROC-AUC and PR-AUC measure are used in this chapter to assess how effective minimal pairs are for detecting allophones. The former measure can also be interpreted as assessing how effective a procedure is for detecting phonemically distinct phones.

Three different algorithms are evaluated: the existence of putative minimal pairs, putative minimal pair counts, and putative minimal pair independent counts. From a theoretical perspective, putative minimal pairs using independent counts should be the preferred algorithm, because counts are not artificially inflated. In fact this algorithm did appear to score marginally better than the others on Kua-nsi.

On TIMIT the general performance of all the minimal pair algorithms was close to chance. The primary reason for this was that known allophones were showing putative minimal pairs. This was due to phonological processes acting across word boundaries, a transcription error, interpretation issues regarding glottalised segments, and a potential morphophonemic process. Other issues such as free variation could cause similar problems.

Pseudo minimal pairs did seem to be the biggest problem for both languages. It might be argued that this was a deficiency with the dataset and which should be cleaned up. However it is important that extra linguistic knowledge is not imposed on the data to be as realistic as possible to the fieldwork survey scenario. There is also the argument that the experimental method should deal with the issue of rare phones, and although this should be investigated in the future, the problem of pseudo minimal pairs is the dominant issue. The poor results in these experiments and the associated problem of pseudo minimal pairs primarily point to a weakness in the theoretical assumptions. They cast doubts about the effectiveness of minimal pairs in a phonemic analysis.

Minimal pairs are often viewed as having a privileged status in establishing contrast. Other procedures in a phonemic analysis are viewed as merely bringing evidence to bear on the question. However as the example of TIMIT shows, it is important to consider all the lines of evidence and not to allow a judgment to be trumped by minimal pairs. The use of the term *putative min-*

imal pair brings the procedure down to the same level as other procedures so that the evidence can be considered together.

Sakel and Everett (2012) argue for a similar approach. They quote Chomsky (1964, p.97) who states "In general it should be observed that 'minimal pair' is not an elementary notion. It cannot be defined in phonetic terms but only in terms of a completed phonemic analysis". However, it is still seen as a useful tool and in referring to Postal (1968) they argue that "We do not thereby eliminate minimal pairs from analyses, but rather we bring the principle of their application into proper perspective".

One practical outcome of this is that the phone relationship chart for minimal pairs is made a more effective tool if a list of minimal pairs can be listed for any potential contrast the linguist wishes to check.

6.7 Chapter summary

The use of minimal pairs is regarded as a particularly effective method in phonemic analysis and the only method to conclusively establish contrast between sounds (Hayes, 2009, p.34). In this chapter minimal pairs are quantitatively evaluated for their effectiveness. The minimal pair algorithms investigated in this chapter have surprisingly poor performance. On TIMIT it is little better than random. There is not much difference in performance between the three variations on the algorithm. From a theoretical perspective, putative minimal pairs using independent counts (MPIC) should be the preferred algorithm. On TIMIT standard counts performed slightly better, but due to this counting method many phone pairs had artificially inflated counts.

In agreement with other authors, the recommendation is not to abandon minimal pair analysis but to consider the results at the same level as other evidence. Using the term putative minimal pair helps to make this perspective clear.

The significant original contributions of this chapter are as follows:

- Three minimal pair algorithms were evaluated on the TIMIT and Kua-nsi corpora
- The independent counts of minimal pairs was introduced
- Analysis of the TIMIT corpus emphasised the importance of the term *putative minimal pair*

Chapter 7

Discussion and Conclusions

The key scientific question underpinning the work in this thesis is *"To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?"*. A secondary question is *"What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?"*

In order to answer these questions the chapter contains the following five sections: a review of the complete phonemic analysis process, a summary of results, the answers to the scientific questions, a summary of original contributions and a section on implications.

7.1 Reviewing the scope of the thesis

A description of the scope of this thesis can be found in the first chapter, Section 1.4. In this current section all the procedures of a phonemic analysis that have not been covered directly in the experiments are considered for their impact. Following this, reasons are given why each procedure was evaluated separately.

7.1.1 Further procedures in phonemic analysis

Figure 7.1 (a duplicate of Figure 1.3) shows how the phonemic analysis process involves much more than has been investigated in this thesis. However it should also be noted that many descriptions in the literature of phonemic analysis narrowly confine the process to just a handful of steps. For example in O'Grady et al. (2005, p.25) the summary flowchart of phonemic analysis involves four yes/no questions regarding minimal pairs, free variation, complementary distribution, and near minimal pairs. Although phonetic similarity is implicit, such a narrowly defined scope of phonemic analysis is very common. The phonemic analysis procedures that are outside the scope of this thesis are considered below.

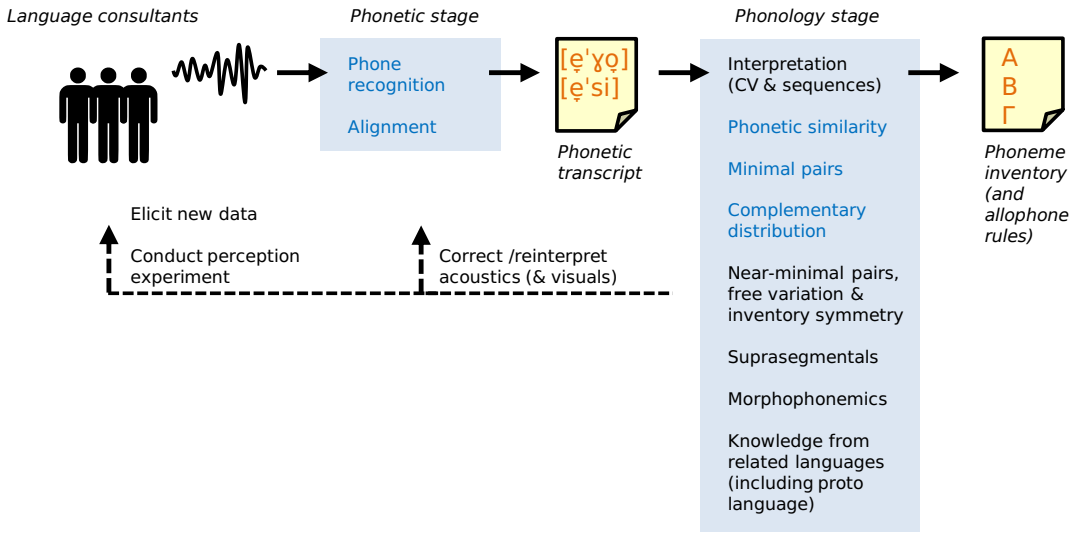

Figure 7.1: The stages in a phonemic analysis. Procedures written in blue (or grey if in monochrome) are those investigated in the thesis.

The interpretation stage, although not investigated fully, was partially addressed in Chapter 5 with the phone transition count matrix. Interpretation brings up some difficult theoretical issues: it is required at the beginning of a phonemic analysis, but there is a need to appeal to phonological generalisations such as syllable structure. These theoretical issues are not addressed in this thesis. If any of the procedures highlight the need for an iterative approach to phonemic analysis, it is the stage of interpretation. It is expected that the phone transition count matrix will be a useful tool to help the linguist in this iterative process: visually highlighting patterns that may become clearer as the cycles of iterations progress.

The investigation into minimal pairs (Chapter 6) has highlighted some theoretical issues that would only be compounded for near-minimal pairs; so it is appropriate that near-minimal pairs are not included in the scope of this thesis.

Of all the issues outside the scope of this thesis, free variation is probably the most relevant to the experiments. When a pair of allophones exhibit free variation there is no complementary distribution because the same environments can occur around both realisations. Free variation can also produce putative minimal pairs that are really pseudo-minimal pairs. In some dialects of US English, tapping can be in free variation (Hayes, 2009, p.60) although, as described in Section 6.4, this does not seem to be the primary reason for the putative minimal pairs found in Experiment 6D. In the phonemic procedure described by Bartram et al. (2008), the linguist is instructed to check the gloss when looking for minimal pairs in case of free variation. This could also be automated e.g. with semantic analysis (Section 6.5).

Inventory symmetry has less concrete influence on the analyses, and it seems that a human is best placed to look for symmetrical patterns of the whole phonological system, although with knowledge of the features, an automated system could highlight patterns that weren't obvious in the standard table layout of a phoneme inventory.

Suprasegmentals can bring additional evidence to clarify ambiguities from a purely segmental phonemic analysis. For example, an apical voiced fricative in Kua-nsi was confirmed to be functioning as a vowel because it was consistently tone-bearing. Influence of features over a suprasegmental timescale such as the dorsal features in vowel harmony could potentially be modelled with factored language models as described in Section 5.7.2.

Morphophonemic analysis brings the ability for suggesting new diagnostic word forms to elicit new data. Morphophonemics can help to expose pseudo-minimal pairs where these are caused by differences in phonological boundaries or displaced contrasts (see Section 6.4).

The use of related languages was helpful in confirming that a putative minimal pair was a pseudo-minimal pair in the discussion after Experiment 6B in Section 6.2. A more general comparison of phonological rules could also be useful.

These procedures that are outside the scope of the experiments in this thesis provide additional evidence for the linguist. Often this is extra information that would clarify existing ambiguities, but sometimes it might also appear to contradict other evidence such as putative minimal pairs. The linguist is expected to look at the evidence from the semi-automated procedures alongside the evidence from these other analysis procedures, and interpret the results together based on their knowledge and experience.

7.1.2 Reasons for separate evaluations

In this thesis each procedure in the phonemic analysis process has been evaluated separately to calculate its individual contribution. Occasionally results have been combined (as in Chapter 5) so that results can be compared against previous studies. The emphasis though has been to evaluate each procedure separately for similar reasons as discussed above. At the current stage of technology, it seems sensible to keep the emphasis on machine-assisted rather than machine-automated phonemic analysis. This means leaving the job to a human expert to weigh up each piece of evidence before combining them and making generalisations. This fits in with the general principle of having as much surface level evidence before positing underlying representations; and postponing abstractions until as much data as possible can be explained. This principle comes across most clearly in Bartram et al. (2008) where the linguist is instructed to check all relevant phone relationships e.g. for minimal pairs and complementary distribution *before* postulating particular rules or phonemes. This has been a guiding principle in this thesis e.g. the phone transition count matrix highlighted a possible vowel neutralisation in Kua-nsi, which could have easily been missed if generalisations had been made too early (see Section

5.1).

With an evaluation measure now established; there is a wealth of further work that could be undertaken in the future to investigate different ways of combining the different procedures. For example, currently most descriptions of a phonemic analysis follow a decision tree flowchart, often starting with a test for minimal pairs (O'Grady et al., 2005, p.25), (Hayes, 2009, p.35). Further experiments may confirm that a question about minimal pairs is not the best first question to ask.

7.2 Summary of results

7.2.1 Standard measures

Figure 7.2 summarises the results from Chapters 3, 5 and 6. The exact values that correspond to these results can be found in the relevant chapters, and also in Appendix A where further precision scores are included. Figure 7.3 summarises the results from Chapter 4.

All the different procedures show a better than chance performance except for the putative minimal pair algorithm when applied to the TIMIT data. This is also the only time that the PR-AUC value falls below the chance level.

The phonetic similarity algorithms investigated in Chapter 3 maintain the same ranking for all three languages. The binary feature edits per phone (BFEPP) algorithm performed best followed by the relative minimal difference (RMD) which was adapted from Peperkamp et al. (2006) to work with binary features. Although the active articulator algorithm (AA) shows a lower performance, it had the advantage of never missing an allophone in the languages tested. The French data was used to make a comparison with previous studies. The French results indicate that TIMIT and Kua-nsi are challenging data, rather than there being a limitation with the binary features system that the phonetic similarity algorithm relies on (see Section 3.4.4 for further information).

The complementary distribution algorithms investigated in Chapter 5 for TIMIT and Kua-nsi are shown in the centre of the bar charts in Figure 7.2. The Jeffreys Divergence (JD) algorithm adapted from Peperkamp et al. (2006) did not make use of features, and performed relatively poorly. The assimilation criterion (AC) also adapted from Peperkamp et al. has a lower performance on TIMIT, this appears to be due to an incompatibility with the feature set used. This led to the development of the assimilating features (AF) algorithm that performed better on both corpora. One result not shown in Figure 7.2 was the successful use of relative entropy to identify the default allophone in an allophone pair. For all the default allophones that were known, the relative entropy algorithm correctly identified them (Section 5.4.2 and 5.6). The phonology visualisations developed in Chapter 5, demonstrate their effectiveness for discovering many phonological patterns. There is much potential for algorithms to take advantage of

identifying these patterns and recognising complementary distribution.

The minimal pair algorithms investigated in Chapter 6 have surprisingly poor performance. On TIMIT it is little better than random. There is not much difference in performance between the three variations on the algorithm. From a theoretical perspective, putative minimal pairs using independent counts (MPIC) should be the preferred algorithm. On TIMIT standard counts performed slightly better, but due to this counting method many phone pairs had artificially inflated counts. In general, compared to the procedures of phonetic similarity and complementary distribution, the minimal pairs procedure performed worst. One striking example is the active articulator algorithm consistently performing better than minimal pairs. This suggests that a knowledge of the active articulators used is more helpful than the use of minimal pairs to determine whether two sounds are phonemically distinct.

The phone recognition and alignment algorithms investigated in Chapter 4 are part of the phonetic stage in the phonemic analysis and are therefore measured differently. This is shown in Figure 7.3. For simplicity and ease of comparison the phone error rate is used in the summary graph to measure recognition. This is a very strict measure, where an error counts as anything that is not an exact match. The same results using the BFEPP phonetic distance measure can be found in Chapter 4. The task of cross-language phone recognition for producing a narrow phonetic transcript is extremely challenging and, similar to a previous study (Walker et al., 2003), there are many errors. Unlike word-based ROVER it was not possible to improve on the best component recogniser error rate for phone-based ROVER. Alignment was evaluated with the standard forced alignment evaluation error; the proportion of boundaries placed more than 20ms away from the manually labelled boundary. Combining the different phone recognisers by choosing the median boundary was shown to improve performance. In using the BFEPP phonetic distance measure to automatically map across phone inventories, any IPA transcript in any language can be aligned with the audio. This allowed a corpus of Bosnian conversation to be automatically aligned using a Russian phone recogniser, halving the time needed when compared to manual alignment (Experiment 4C, Section 4.3.2).

7.2.2 Algorithm complexity

Table 7.1 shows the time complexity of the algorithms used in this thesis. Both the active articulator and BFEPP algorithm take each phone and compare it against every other phone, so the time complexity is written as $O(p^2)$ where p is the number of unique phones. Peperkamp's relative minimal difference algorithm has an additional level where each phone pair is then again compared with every phone in the set so the time complexity is $O(p^3)$. Since there are many unique sounds in a narrow transcription, algorithms with time complexity $O(p^3)$ can take longer to execute than word-based algorithms with complexity $O(w^2)$ where w is the number of unique words. The duration of the recorded audio can also be significant. For example

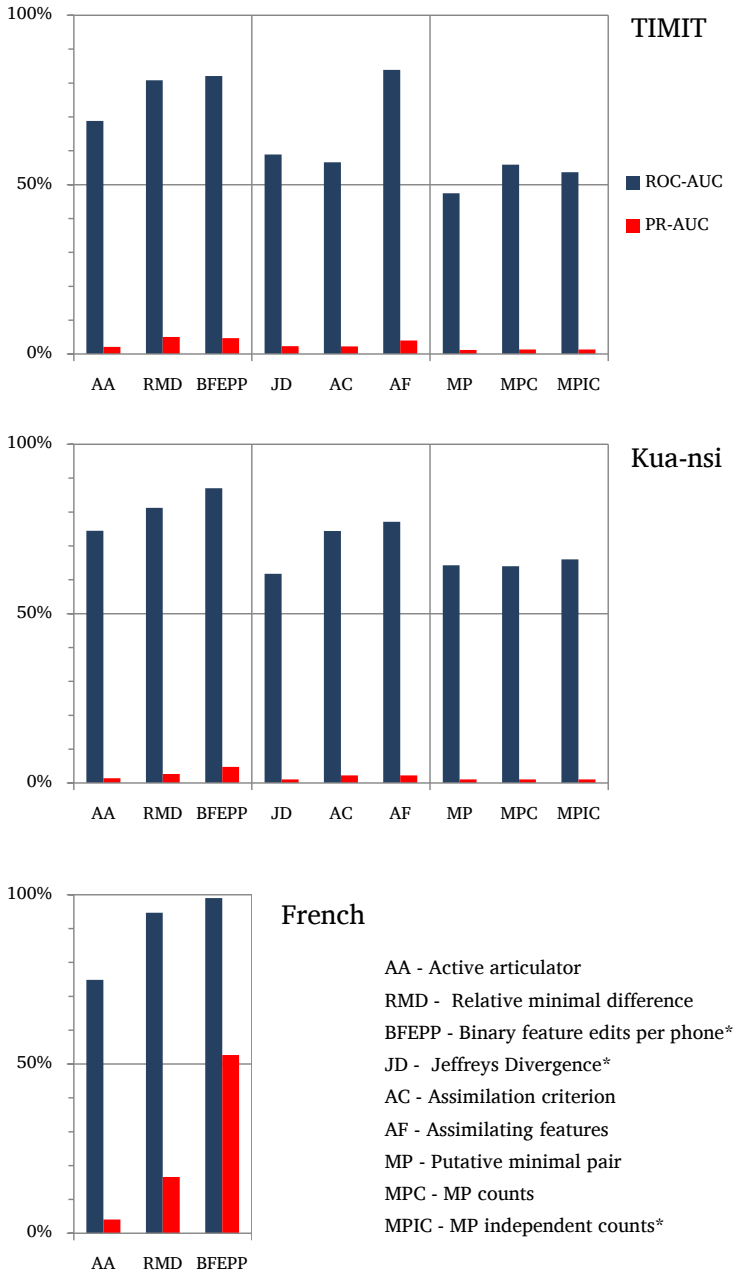

Figure 7.2: Graph showing summary results of phonemic analysis procedures. The vertical lines indicate the groupings of algorithms into the phonetic distance, complementary distribution and minimal pair procedures. The horizontal line at 50% indicates the chance level for the ROC-AUC values (the chance values for PR-AUC are not shown). *Algorithms with an asterisk are those that best represent each procedure.

Figure 7.3: Graph showing summary results of the phone recognition and alignment procedures. The Y-axis is drawn top-down to highlight that error is being measured, although equivalent accuracy can be inferred from the white bars. Note that the median combination method on the TIMIT alignment shows a statistically significant improvement over the component recognisers.

Algorithm	Time complexity
Active articulator	$O(p^2)$
Relative minimal difference	$O(p^3)$
Binary Feature Edits Per Phone (BFEPP)	$O(p^2)$
Jeffreys Divergence	$O(p^3)$
Assimilation criterion	$O(p^3)$
Assimilating features	$O(p^3)$
Putative minimal pair (MP)	$O(w^2)$
MP counts	$O(w^2)$
MP independent counts	$O(w^2)$
Phone recognition or alignment	O(t)

Table 7.1: Time complexity of the different algorithms; p = number of unique phones, w = number of unique words, t = time length of recording

in Kua-nsi $p > 100$ including tones, $w = 500$, and time $t = 1$ hour. This was particularly noticeable when each fundamental operation was computationally expensive such as Jeffreys Divergence. Most algorithms took seconds or minutes to complete, whereas Jeffreys Divergence and phone recognition/alignment could take a matter of hours. The new algorithms introduced in this thesis were either equal in complexity or were less complex than Peperkamp's algorithms. BFEPP is an example of a very simple algorithm where only the relevant two phones need to be used as the input.

7.2.3 Explicit efficiency savings

In Chapter 1, there was a reference to a statement by Hockett that it takes an experienced linguist 10 days of hard work to complete 90% of an analysis. It was stated that although there is much variability in this figure there is consensus from contemporary linguists that this is still a good rule of thumb. The evaluation measures ROC-AUC and PR-AUC have intuitive interpretations (see Section 3.2.1), but the time saving for a linguist is also an intuitive measure. Although Hockett (1955) does not elaborate precisely on what a 90% analysis means, from reading the rest of the article it is not unreasonable to suggest that this equates to a discovery of 90% of the allophonic relationships. In theory this would involve the linguist working through all the different phone pairs, e.g. checking for evidence of contrast or for an allophonic relationship. Of course there would be shortcuts, e.g. implicit rejection of phonetically dissimilar sounds, but with no outside assistance the linguist would need to work through on average 90% of the phone pairs to discover 90% of the allophonic relationships. However, if the list of phone pairs was ordered in a way so that allophone pairs were more likely to be promoted to the top of the list, then it might be possible that the linguist only needs to work through 45% of the phone

Algorithm applied to Kua-nsi	Fraction of time
Active articulator	51%
Relative minimal difference	26%
Binary Feature Edits Per Phone (BFEPP)	21%
Jeffreys Divergence	80%
Assimilation criterion	71%
Assimilating features	64%
Putative minimal pair (MP)	66%
MP counts	66%
MP independent counts	66%

Table 7.2: Time savings for the linguist: fraction of time needed to discover all but one of the consonant allophone pairs in the Kua-nsi corpus when compared to a randomly ordered list of pairs (smaller values indicate a better performing algorithm e.g. BFEPP gives the best time saving)

pairs to discover 90% of the allophone pairs, thereby halving the time needed.

The closest to a 90% proportion of allophone pairs is 83% in the Kua-nsi corpus, equating to all but one of the consonant allophone pairs. Fortunately it is possible to take the full description in Hockett (1955) and make an interpolation[1] that 83% of the analysis should take 6.7 days.

Table 7.2 shows the time savings expressed as the fraction of work or time needed to discover 83% of the consonant allophone pairs in the Kua-nsi corpus compared to a randomly ordered list of pairs. The values have been calculated from the 83% recall point on the ROC curves. This time saving assumes that the only change to the linguist's workflow is to present them with a sorted list of phone pairs at the beginning of their manual analysis based on the appropriate algorithm. For example, using the BFEPP algorithm, the fraction of time is 21% i.e. reducing the time from 6.7 days to 1.4 days. Obviously many simple assumptions are made, partly because it is difficult to characterise the shortcuts a linguist takes in an analysis and how this is affected by presenting them with a list. However, the speed-up could improve with the interactive use of the tools; e.g. using phone relationship charts and phone transition count matrices available as the linguist follows through the iterative cycles of interpretation and analysis. Therefore the values given for estimated time savings are no substitute for the more comprehensive ROC-AUC and PR-AUC scores, but they do help in providing an intuition of the benefit in using the algorithms.

[1] The description given by Hockett (1955) can be modelled by the formula $y = 1 - \frac{1}{kx}$ where k = 0.9, x is the number of days, and y is the proportion of the analysis completed

7.3 Answers to scientific questions

With the results summarised, it is now possible to answer the scientific questions. The first question is *"To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?"*. A very basic answer is that a machine algorithm can contribute by performing with an accuracy that is better than chance. This is true for all the procedures investigated in the phonology stage. This can be seen in Figure 7.2 by all the ROC-AUC scores that are above the 50% line. The ROC-AUC evaluation measure particularly with its probabilistic interpretation, demonstrates that there is a measurable contribution from each algorithm. The explicit efficiency savings shown in Table 7.2 give an additional intuitive measure of the benefit of each algorithm.

At the phonetic stage, cross-language phone recognition had too many errors to be practically beneficial. However cross-language forced alignment has been shown to take a tenth of the time needed when compared to manual phone alignment. (Experiment 4C, Section 4.3.2).

The secondary scientific question is *"What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?"*

For each of the procedures there is a principal algorithm that represents each procedure best. For the main two datasets TIMIT and Kua-nsi, the best phonetic similarity algorithm, BFEPP resulted in an average ROC-AUC of 85%. The primary complementary distribution algorithm, Jefferys Divergence resulted in an average ROC-AUC of 60%. Although strictly not a pure complementary distribution algorithm, assimilating features which gave an average ROC-AUC of 81%, indicates the importance of considering features. The primary minimal pairs algorithm, using independent counts resulted in an average ROC-AUC of 60%.

Given the best available data and the machine-assisted procedures described, the results give a strong indication that phonetic similarity is the most important piece of evidence in a phonemic analysis. It is also a fundamental part of the phone alignment algorithm.

The complementary distribution algorithm appears to have potential for improvement; the use of phonological features, such as binary features, is the most promising area.

As described above, it can be seen that minimal pairs contributed very little on their own. On investigating the reasons behind this, it was recommended that in a phonemic analysis they are referred to as putative minimal pairs. The experiments have underlined the importance of keeping the human in the loop i.e. it is machine-assisted phonemic analysis not machine-automated phonemic analysis.

7.4 Contributions

A summary of each chapter's primary original contributions are as follows:
Chapter 3: Phonetic similarity

- Three phonetic similarity algorithms were evaluated on the TIMIT and Kua-nsi corpora
- Statistical measures have been applied to quantitatively evaluate phonemic analysis
- The data used in the experiments is more phonetically accurate than previous studies
- The active articulator algorithm was introduced for predicting phonemically distinct phones
- The BFEPP algorithm was introduced as the best performing phonetic similarity measure

Chapter 4: Phone recognition and alignment

- Cross-language phone recognition and alignment was evaluated on the TIMIT and Bosnian corpora
- A phone-based rather than word-based ROVER voting system was introduced
- Cross-language forced alignment was introduced with automatic IPA mapping (using BFEPP)
- Combining different language phone recognisers was shown to improve alignment results

Chapter 5: Complementary distribution

- Three complementary distribution related algorithms were evaluated on the TIMIT and Kua-nsi corpora
- The phone transition count matrix was introduced to visualise the phonology of a language
- Peperkamp's assimilation algorithm was adapted for Hayes' feature set and improved

Chapter 6: Minimal pairs

- Three minimal pair algorithms were evaluated on the TIMIT and Kua-nsi corpora
- The independent counts of minimal pairs was introduced
- Analysis of the TIMIT corpus emphasised the importance of the term *putative minimal pair*

There have also been some practical contributions of the work including tangible benefits to the analysis of two under-resourced languages. The consonants in the Kua-nsi dataset have been used in the evaluation of the different algorithms but, as stated in Section 1.4, there was some uncertainty about the vowel system. Many of the algorithms were still applied to the vowels; the findings, particularly from using the phone transition count matrix, were then passed back to the field linguists. This influenced the decisions made about the orthography and alphabet[2], e.g. it was decided that the apical vowel [ɿ] would not be represented by a separate symbol because it was the allophone of another underlying form that was already represented in the alphabet.

[2]Crook and Castro (2012), personal communication

The cross-language forced alignment on Bosnian has helped to create the four channel 3.8 hour corpus described in Kurtic et al. (2012), a dataset that is currently being used in conversational analysis research.

The primary contribution of this thesis is the answer to the scientific questions. It is shown at the machine-assisted approach can help in phonemic analysis. The important role of phonetic similarity, the potential of feature-based complementary distribution algorithms, and the limitation of minimal pairs have been highlighted.

7.5 Implications

Some of the tools and techniques described in this thesis are ready to be used now to assist linguists in conducting a phonemic analysis; these are described below. There are also implications for what to prioritise for future research. It is expected that more suitable data will become available over time as archiving of primary documentation becomes more routine for linguists and corresponding phonological analyses of the languages become more mature (Section 3.4.2). The current priority is to evaluate the algorithms on both consonants and vowels. Two areas of research that need more investment but have potential for giving good returns are phone-based ROVER (Section 4.2.3) and a probabilistic feature-based complementary distribution algorithm (Section 5.7.2). Also of high priority is the comparison of the BFEPP phonetic distance measure with the phonetic distance measures in Mielke (2009) (see Section 3.2.3) e.g. to investigate correlations.

There are also some implications for theoretical phonology. In summarising the historical development of phonemic analysis, it was seen in Section 2.2 that there was originally much optimism in automating the process until it became clear that the process was less deterministic than first thought. Rather than steps of self contained analysis, phonemic analysis rightly became reliant on iterative cycles spanning the whole process and bringing in many knowledge sources. The experiments in this thesis have shown that a machine can help in a broad range of procedures in phonemic analysis, not just parts of the analysis that would be traditionally viewed as mechanical or deterministic. If optimism doesn't return for automating phonemic analysis, there should be at least some optimism for machine-assisted phonemic analysis.

As explained in Section 6.1.2, the ROC-AUC statistic used in this thesis not only measures the effectiveness of each algorithm in detecting allophones but simultaneously measures the effectiveness of each algorithm in detecting phonemically distinct phones. It is interesting that not only do non-minimal-pair methods work well in detecting phonemically distinct phones, but that the use of minimal pairs is less effective. This finding appears to be in disparity with the claim that "by far the most effective method in phonemicization is to look for minimal pairs" (Hayes, 2009, p.34). It is possible that this statement implicitly included the use of phonetic

similarity, and the effectiveness for phonemicization[3] is not specifically defined. However the findings in this thesis do cast doubt on any premise that minimal pairs alone are the most effective method for detecting phonemically distinct phones in a phonemic analysis.

The most important implications are practical. As described above, there have been the tangible benefits to two under-resourced languages. Also a number of linguists working with endangered languages have shown an interest in the work of this thesis. It is expected that the tools and techniques developed in this thesis, will be used to help document further under-resourced languages. Currently the most mature tools that can be used immediately are cross-language forced alignment for aligning any IPA transcription in any language to the audio and the phone count transition matrix for visualising the phonology of any language. To this end it is the intention to release software, utilities, and documentation so others can make free use of these resources. This is progressively being made available on the internet.[4] It is hoped that many more languages will benefit from undergoing a phonemic analysis, leading to further linguistic analysis, literacy, and access to speech technology.

7.6 Practical outcomes for the field linguist

The practical implications are of most value to the field linguist. If a linguist is at the phonetic stage of an analysis, they may want to conduct an acoustic analysis e.g. measure the vowel formant space, the duration of certain phones such as stop consonants, or different tones. Normally this would require a lengthy procedure of manually aligning labels with the acoustic data. Cross-language forced alignment allows this to be done automatically. The linguist just needs to provide an audio file for each utterance, and a text file of each IPA transcription. The output from the algorithm is the time aligned phone transcript in HTK or Praat format e.g. see Figure 4.3 (p.76). There is other software (Lennes, 2011) that can make use of these time alignments for acoustic analysis such as vowel F1 F2 plots.

If a linguist is at the phonological stage of analysis, the phone transition count matrix can instantly give an initial visualisation of the language's phonology. The linguist just needs to provide the IPA wordlist, and the script will produce the matrix e.g. Figure 5.2 (p.89) was created from the wordlist in Castro et al. (2010).

In conducting a phonemic analysis, a linguist will be interested to know if certain pairs of phones are phonemically distinct or are in an allophonic relationship. The phonetic similarity, complementary distribution, minimal pair algorithms can all be used to give suggestions for this. The linguist just needs to provide the IPA wordlist and these algorithms will provide a ranked list of phone pairs indicating which pairs are most likely to be allophones of the same

[3] "Phonemicization is the body of knowledge and techniques that can be used to work out the phonemic system of a language" (Hayes, 2009, p.34)

[4] http://speechchemistry.wordpress.com

phoneme. This list can also be displayed as a phone relationship chart e.g. Figure 3.4 (p.51). The time savings in using these algorithms are estimated above in Section 7.2.3. As discussed earlier the linguistic should use their knowledge and experience in interpreting the output. Even if a linguist decides not to use these algorithms, the findings showing the relative merits of each procedure in phonemic analysis are of value. In particular, the field linguist should be wary of giving too much weight to minimal pairs and should consider them alongside the evidence from other procedures in a phonemic analysis.

7.7 Chapter summary

The key scientific question underpinning the work in this thesis is *"To what extent can a machine algorithm contribute to the procedures needed for a phonemic analysis?"*. A secondary question is *"What insights does such a quantitative evaluation give about the contribution of each of these procedures to a phonemic analysis?"*

It is demonstrated that a machine-assisted approach can make a measurable contribution to a phonemic analysis for all the procedures investigated; phonetic similarity, phone recognition & alignment, complementary distribution, and minimal pairs. The evaluation measures introduced in this thesis allows a comprehensive quantitative comparison between these phonemic analysis procedures. Given the best available data and the machine-assisted procedures described, there is a strong indication that phonetic similarity is the most important piece of evidence in a phonemic analysis. It is also a fundamental part of the phonetic alignment algorithm. Featured-based complementary distribution algorithms are shown to have much potential for improvement, and the limitations of minimal pairs have been highlighted.

The tools and techniques developed in this thesis have resulted in tangible benefits to the analysis of two under-resourced languages, Kua-nsi and Bosnian, and it is expected that many more languages will follow.

References

Aimetti, G., R. K. Moore, and L. ten Bosch (2010). Discovering an Optimal Set of Minimally Contrasting Acoustic Speech Units: A Point of Focus for Whole-Word Pattern Matching. In *Proc. Interspeech*.

Aimetti, G., R. K. Moore, L. ten Bosch, O. J. Räsänen, and U. K. Laine (2009). Discovering keywords from cross-modal input: Ecological vs. engineering methods for enhancing acoustic repetitions. In *Proc. Interspeech*.

Alomofor, C. and S. C. Anderson (2005). Awing Orthography Guide, December 2005 Revision. *SIL Cameroon*.

Arlotto, A. (1981). *Introduction to historical linguistics*. University Press of America.

Aslam, J. and E. Yilmaz (2005). A geometric interpretation and analysis of R-precision. In *Proc. 14th ACM international conference on Information and knowledge management*, pp. 664–671. ACM.

Atkinson, Q. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science 332*(6027), 346.

Bacchiani, M. and M. Ostendorf (1998). Using automatically-derived acoustic sub-word units in large vocabulary speech recognition. In *Fifth International Conference on Spoken Language Processing*. ISCA.

Bacchiani, M. and M. Ostendorf (1999). Joint lexicon, acoustic unit inventory and model design. *Speech Communication 29*(2-4), 99–114.

Bagwell, C. et al. (2009). SoX - Sound eXchange. *sox-14.3.0*. [Software].

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology 12*(4), 387–415.

Bartram, C. (Ed.) (2008). *Introduction to Phonology (Full phonemic procedure section)*. European Training Programme (UK Campus).

Berkling, K. (1996). *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*. Ph. D. thesis, Syracuse University.

Bevan, D. (1995). *FindPhone Users's Guide: Phonological Analysis for the Field Linguist; Version 6.0*. SIL.

Bird, S. and G. Simons (2003). Seven dimensions of portability for language documentation and description. *Language*, 557–582.

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language 24*(1), 3–46.

Boersma, P., P. Escudero, and R. Hayes (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In *Proc.15th International Congress of Phonetic Sciences*, Volume 1013, pp. 1016.

Boersma, P. and D. Weenink (2011). Praat: doing phonetics by computer. *Version 5*, 2. [Software].

Brenzinger, M., A. Yamamoto, N. Aikawa, D. Koundiouba, A. Minasyan, A. Dwyer, C. Grinevald, M. Krauss, O. Miyaoka, O. Sakiyama, et al. (2003). Language vitality and endangerment. Paris: UNESCO Expert Meeting on Safeguarding Endangered Languages.

Buckley, C. et al. (2006). The Trec_eval evaluation package v8.1. *NIST*. [Software].

Burget, L., P. Matějka, and J. Černocký (2006). Discriminative training techniques for acoustic language identification. *Proc. ICASSP*.

Burget, L., P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, et al. (2010). Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models. *Proc. ICASSP*.

Burquest, D. (2006). *Phonological analysis: a functional approach*. SIL International.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press.

Campbell, W., J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language 20*(2-3), 210–229.

Casali, R. (2009). Dekereke: A software tool for phonological fieldwork. In *6th World Congress of African Linguistics*.

Castro, A. and G. Chaowen (2010). Phonological innovation among Hmong dialects of Wenshan. *Journal of the Southeast Asian Linguistics Society 3*(1), 1–39.

Castro, A., B. Crook, and R. Flaming (2010). A sociolinguistic survey of Kua-nsi and related Yi varieties in Heqing county, Yunnan province, China. *SIL Electronic Survey Reports 1*, 96.

Chao, Y. (1930). ə sɪstəm əv —toun-lɛtəz. *Le Maitre Phonetique 30*, 24–27.

Chollet, G., J. Cernocky, A. Constantinescu, S. Deligne, and F. Bimbot (1999). Toward ALISP: A proposal for automatic language independent speech processing. *NATO ASI series. Series F: computer and system sciences*, 375–388.

Chomsky, N. (1964). *Current issues in linguistic theory*. Number 38. Mouton.

Clark, J., C. Yallop, and J. Fletcher (2007). *An Introduction to Phonetics and Phonology*. Blackwell Publishing.

Crystal, D. (2000). *Language death*. Cambridge Univ Press.

Daniels, P. and W. Bright (1996). *The world's writing systems*. Oxford University Press.

Davis, J. and M. Goadrich (2006). The relationship between Precision-Recall and ROC curves. In *Proc. 23rd International Conference on Machine Learning*, pp. 233–240. ACM.

De Wachter, M., M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle (2007). Template-based continuous speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 15*(4), 1377 –1390.

Demuth, K. (2007). Sesotho Speech Acquisition. *The international guide to speech acquisition*, 528–538.

Deng, L. and D. Sun (1994). A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *Journal of the Acoustical Society of America 95*(5), 2702–2719.

Dingemanse, M. (2008). Review of Phonology Assistant 3.0.1. *Language Documentation & Conservation 2*(2), 325–331.

Dollaghan, C., M. Biber, and T. Campbell (1993). Constituent syllable effects in a nonsense-word repetition task. *Journal of Speech and Hearing Research 36*(5), 1051.

Dryer, M. and M. Haspelmath (2011). The world atlas of language structures online. *Munich: Max Planck Digital Library*.

Duanmu, S. (2000). *The phonology of standard Chinese*. Oxford University Press.

Eimas, P., J. Miller, and P. Jusczyk (1987). On infant speech perception and the acquisition of language. *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press.

Eimas, P., E. Siqueland, P. Jusczyk, and J. Vigorito (1971). Speech Perception in Infants. *Science 171*(3968), 303.

El Hannani, A. (2007). *Text-independent speaker verification based on high-level information extracted with data-driven methods*. Ph. D. thesis.

Esling, J. et al. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Feldman, N. and T. Griffiths (2007). A rational account of the perceptual magnet effect. In *Proc. 29th Annual Conference of the Cognitive Science Society*, pp. 257–262.

Fennell, C. and J. Werker (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech 46*(2-3), 245–264.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*.

Fitt, S. and S. Isard (1999). Synthesis of regional English using a keyword lexicon. In *Sixth European Conference on Speech Communication and Technology*.

Fromkin, V. and R. Rodman (1998). *An introduction to language (Sixth edition)*. Holt, Rinehart and Winston.

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. 20th International Joint Conference on Artificial Intelligence*, Volume 6, pp. 12. Morgan Kaufmann Publishers Inc.

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue (1993). TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Gauvain, J., A. Messaoudi, and H. Schwenk (2004). Language recognition using phone lattices. *Proc. ICSLP*.

Gildea, D. and D. Jurafsky (1996). Learning Bias and Phonological-Rule Induction. *Computational Linguistics 22*(4), 497–530.

Gleason, H. (1961). *An introduction to descriptive linguistics*. Holt, Rinehart and Winston, Inc.

Goldman, J. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. *Proc. Interspeech*.

Goldsmith, J. (1976). *Autosegmental phonology*. Massachusetts Institute of Technology.

Gopnik, A., A. Meltzoff, and P. Kuhl (1999). *The Scientist in the Crib: Minds, Brains, and How Children Learn*. William Morrow & Col.

Greenberg, S., J. Hollenback, and D. Ellis (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proc. ICSLP*.

Grenoble, L. and L. Whaley (2006). *Saving languages: An introduction to language revitalization*. Cambridge University Press.

Gussenhoven, C. and H. Jacobs (2005). *Understanding phonology*. Hodder Arnold.

Halle, M. (1959). *The Sound Pattern of Russian*. Mouton.

Harnad, S. (2003). Categorical Perception (in Encyclopedia of Cognitive Science). Nature Publishing Group: Macmillan.

Harris, Z. (1951). *Methods in structural linguistics*. University of Chicago Press.

Hassan, S. and R. Mihalcea (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, pp. 1192–1201. Association for Computational Linguistics.

Hayes, B. (2009). *Introductory phonology*. Wiley-Blackwell.

Hazen, T. and V. Zue (1993). Automatic Language Identification using a Segment-based Approach. *Proc. Eurospeech*, 1303–1306.

Heeringa, W., P. Kleiweg, C. Gooskens, and J. Nerbonne (2006). Evaluation of string distance algorithms for dialectology. *Linguistic Distances Workshop*, 51–62.

Hieronymus, J. (1993). ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetic Association 23*.

Hieronymus, J., M. Alexander, C. Bennett, I. Cohen, D. Davies, J. Dalby, J. Laver, W. Barry, A. Fourcin, and J. Wells (1990). Proposed speech segmentation criteria for the SCRIBE project. *SCRIBE Project Report*.

Himmelmann, N. et al. (2002). Documentary and Descriptive Linguistics (full version). *Lectures on endangered languages 5*, 37–83.

Hockett, C. (1955, May). How to learn Martian. *Astounding Science Fiction*, 97–102.

Hosom, J. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication 51*(4), 352–368.

House, A. and E. Neuburg (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *The Journal of the Acoustical Society of America 62*, 708.

Huckvale, M. (2004). ACCDIST: a metric for comparing speakers' accents. In *Proc. ICSLP*.

Huckvale, M., R. K. Moore, RSRE, NPL, CUED, CSTR, and UCL (1989). SCRIBE - Spoken Corpus of British English.

Hunt, G. (2008). A comparison of phonology tools. In *SIL Forum for Language Fieldwork*, Volume 9, pp. 6.

Jančík, Z., O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karafiát, P. Matějka, T. Mikolov, A. Strasheim, and J. Černocký (2010). Data selection and calibration issues in automatic language recognition–investigation with BUT-AGNITIO NIST LRE 2009 system. *Proc. Odyssey*.

Jeffreys, H. (1948). *Theory of probability* (2 ed.). Oxford University Press.

Johnson, E. and P. Jusczyk (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language 44*(4), 548–567.

Jurafsky, D. and J. H. Martin (2008, May). *Speech and Language Processing* (2 ed.). Prentice Hall.

Jusczyk, P., P. Luce, and J. Charles-Luce (1994). Infants'sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language 33*(5), 630–645.

Kempton, T. (2007). Language characterisation in speech. *Unpublished PhD thesis transfer report*.

Kempton, T. and R. Moore (2008). Language Identification: Insights from the Classification of Hand Annotated Phone Transcripts. *Proc. Odyssey, Stellenbosch, South Africa*.

Kempton, T., R. Moore, and T. Hain (2011). Cross-language phone recognition when the target language phoneme inventory is not known. *Proc. Interspeech, Florence, Italy*.

Kempton, T. and R. K. Moore (2009). Finding Allophones: an Evaluation on Consonants in the TIMIT Corpus. *Proc. Interspeech, Brighton, UK*, 1651–1654.

Kenny, P., G. Boulianne, P. Ouellet, and P. Dumouchel (2007). Joint factor analysis versus eigenchannels in speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 15*(4), 1435–1447.

Kirchhoff, K. (1998). Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. *Proc. ICSLP 98*.

Kirchhoff, K., J. Bilmes, and K. Duh (2008). Factored language models tutorial. *Technical Report, University of Washington, Seattle UWEETR-2008-0004*.

Kirchner, R., R. Moore, and T. Chen (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics 38*(4), 540–547.

Kohler, J. (1998). Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. *Proc. ICASSP*.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities 37*(3), 273–291.

Kuhl, P. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *The Journal of the Acoustical Society of America 70*, 340.

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Percept Psychophys 50*(2), 93–107.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience 5*(11), 831–843.

Kuhl, P. and J. Miller (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America 63*, 905.

Kullback, S. and R. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Kurtic, E., B. Wells, G. J. Brown, T. Kempton, and A. Aker (2012). A Corpus of Spontaneous Multi-party Conversation in Bosnian Serbo-Croatian and British English. *International Conference on Language Resources and Evaluation, Instanbul, Turkey*.

Labov, W., S. Ash, and C. Boberg (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*, Volume 1. Mouton de Gruyter.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.

Ladefoged, P. and K. Johnson (2010). *A course in phonetics (6th edition)*. Wadsworth Pub Co.

Lamel, L., J. Gauvain, and O. LIMSI-CNRS (1993). Cross-lingual experiments with phone recognition. *Proc. ICASSP 2*.

Lamel, L., J. Gauvain, and O. LIMSI-CNRS (1994). Language identification using phone-based acoustic likelihoods. *Proc. ICASSP 1*.

Lasky, R., A. Syrdal-Lasky, and R. Klein (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *J Exp Child Psychol 20*(2), 215–25.

Le Calvez, R., S. Peperkamp, and E. Dupoux (2007). Bottom-up learning of phonemes: A computational study. In *Proc. Second European Cognitive Science Conference*, pp. 167–172.

Lehonkoski, R., Z. H. Wei, and B. Crook (2010). Simple phonology sketch of Kua-nsi spoken in Heqing. *Unpublished document*.

Lei, X., W. Wu, W. Wang, A. Mandal, and A. Stolcke (2009). Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversation. In *Proc. Interspeech*.

Lennes, M. (2011). collect_formant_data_from_files.praat v.4.0.23. *SpeCT - The Speech Corpus Toolkit for Praat*. [Software].

Lewis, M. (2009). *Ethnologue: Languages of the world*. SIL International.

Lewis, M. and G. Simons (2010). Assessing Endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique/Romanian Review of Linguistics 2*.

Lewis, P. M. (2006). Towards a Categorization of Endangerment of the World's Languages. *SIL Electronic Working Papers*.

Lööf, J., C. Gollan, and H. Ney (2009). Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system. In *Proc. Interspeech*.

Li, H., B. Ma, and C.-H. Lee (2007). A Vector Space Modeling Approach to Spoken Language Identification. *Audio, Speech and Language Processing, IEEE Transactions on 15*, 271–284.

Lin, Y. (2005). *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition*. Ph. D. thesis, University of California.

Lisker, L. and A. Abramson (1970). The voicing dimension: Some experiments in comparative phonetics. *Proc. Sixth International Congress of Phonetic Sciences*, 563–567.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 1198–1202.

Lloyd-Thomas, H., E. Parris, and J. Wright (1998). Recurrent Substrings and Data Fusion for Language Recognition. *Proc. ICSLP, Sydney, Australia, Dec*.

Lyu, D., S. Siniscalchi, T. Kim, and C. Lee (2008). Continuous Phone Recognition without Target Language Training Data. *Proc. Interspeech, Brisbane, Australia*.

Maddieson, I. (1984). *Patterns of Sound*. Cambridge.

Marlett, S. (2005). A typological overview of the Seri language. *Linguistic Discovery 3*(1), 54–73.

Marlett, S., F. Herrera, and G. Astorga (2005). Illustrations of the IPA: Seri. *Journal of the International Phonetic Association 35*(1), 117–121.

Martin, T., B. Baker, E. Wong, and S. Sridharan (2006). A syllable-scale framework for language identification. *Computer Speech & Language 20*(2-3), 276–302.

Matějka, P., P. Schwarz, L. Burget, and J. Černocký (2006). Use of Anti-models to further improve state-of-the-art PRLM Language Recognition Systems. *Proc. ICASSP-06 1*, 197–200.

Matějka, P., P. Schwarz, J. Černocký, and P. Chytil (2005). Phonotactic Language Identification using High Quality Phoneme Recognition. *Proc. Eurospeech*, 2237–2240.

Maye, J., J. Werker, and L. Gerken (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition 82*(3), B101–B111.

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press, USA.

Mielke, J. (2009). A phonetically-based phonetic similarity metric. NELS.

Moore, R. (2007). Spoken language processing: Piecing together the puzzle. *Speech communication 49*(5), 418–435.

Moore, R., M. Russell, and M. Tomlinson (1983). The discriminative network: a mechanism for focusing recognition in whole-word pattern matching. In *Proc. ICASSP*, Volume 8, pp. 1041–1044. IEEE.

Moseley, C. (2009). Atlas of the world's languages in danger. *Paris: UNESCO 13*, 2009.

Muthusamy, Y. (1993). *A segmental approach to automatic language identification*. Ph. D. thesis, Oregon Graduate Institute of Science and Technology.

Muthusamy, Y., R. Cole, and B. Oshika (1992). The OGI multi-language telephone speech corpus. *Proc. ICSLP*, 895–898.

Nakagawa, S., Y. Ueda, and T. Seino (1992). Speaker-independent, text-independent language identification by HMM. *Proc. ICSLP*, 1011–1014.

Nerbonne, J. and W. Heeringa (2010). Measuring dialect differences. In *Theories and Methods Vol. in series Language and Space*, Chapter Measuring dialect differences. Mouton de Gruyter.

NIST (2009). Sclite v2.4 score speech recognition system output. [Software].

Odden, D. (2005). *Introducing phonology*, Chapter Online Chapter 3 extension: Phonemes, contrasts, and phonetic detail. Cambridge University Press.

O'Grady, W., J. Archibald, and R. Kirchner (2005). *Contemporary Linguistic Analysis: An Introduction - Alberta Edition*. Pearson.

Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string'model of speech. *Proc. IEEE ASRU Workshop*.

Ostler, N. et al. (2008). Foundation for Endangered Languages Manifesto.

Oxford (2010). *Oxford Dictionaries*. Oxford University Press.

Paliwal, K. (1990, apr). Lexicon-building methods for an acoustic sub-word based speech recognizer. In *Proc. ICASSP*, pp. 729 –732 vol.2.

Parandekar, S. and K. Kirchhoff (2003). Multi-stream language identification using data-driven dependency selection. *Proc. ICASSP 1*.

Park, A. and J. Glass (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio Speech and Language Processing 16*(1), 186.

Parker, S. (2010). Chamicuro data: exhaustive list (Datos del chamicuro: lista exhaustiva). *SIL Language and Culture Documentation and Description 12*, 59.

Pearce, M. (2007). *The interaction of tone with voicing and foot structure: Evidence from Kera phonetics and phonology*. Ph. D. thesis.

Peddinti, V. and K. Prahallad (2011). Exploiting phone-class specific landmarks for refinement of segment boundaries in TTS databases. *Proc. ICASSP*.

Pegg, J. and J. Werker (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America 102*(6), 3742–3753.

Peperkamp, S., R. Le Calvez, J. Nadal, and E. Dupoux (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition 101*(3), B31–B41.

Petrovska-Delacrétaz, D., M. Abalo, A. Hannani, and G. Chollet (2003). Data-driven Speech Segmentation for Language Identification and Speaker Verification. *ISCA Tutorial and Research Workshop on Non-Linear Speech Processing*.

Pike, K. (1947). *Phonemics, a technique for reducing languages to writing*. University of Michigan Press.

Poser, B. (2008). Minpair. *Version 5.1*. [Software].

Postal, P. (1968). *Aspects of phonological theory*. Harper & Row New York.

Riek, L., W. Mistretta, and D. Morgan (1991). Experiments in language identification. Technical report, Technical Report SPCOT-91-002, Lockheed Sanders, Inc., Nashua, NH.

Saffran, J. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science 12*(4), 110–114.

Sakel, J. and D. L. Everett (2012). *Linguistic fieldwork: A student guide*. Cambridge.

Sapir, E. (1925). Sound patterns in language. *Language*, 37–51.

Schultz, T. and K. Kirchhoff (2006). *Multilingual Speech Processing*. Academic Press.

Schultz, T., I. Rogina, and A. Waibel (1996). LVCSR-based language identification. *Proc. ICASSP*.

Schultz, T. and A. Waibel (1998). Multilingual and Crosslingual Speech Recognition. *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 259–262.

Schwarz, P. (2009). *Phoneme recognition based on long temporal context*. Ph. D. thesis.

Schwarz, P., P. Matějka, L. Burget, and O. Glembek (2009). Phoneme recognition based on long temporal context. *phnrec v2.21*. [Software].

Shriberg, L., D. Austin, B. Lewis, J. McSweeny, and D. Wilson (1997). The percentage of consonants correct (PCC) metric: extensions and reliability data. *Journal of Speech, Language, and Hearing Research 40*(4), 708.

SIL (2007). Speech Analyzer v.3.01. *SIL International*. [Software].

SIL (2008). Phonology Assistant v3.0.1. *SIL International*. [Software].

SIL (2010). Phonology Template Editor and Search Tool (PTEST). *SIL International*. [Software].

Singer, E., P. Torres-Carrasquillo, T. Gleason, W. Campbell, and D. Reynolds (2003). Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification. *Proc. Eurospeech 3*, 1345–1348.

Siniscalchi, S., J. Reed, T. Svendsen, and C. Lee (2010). Exploiting Context-Dependency and Acoustic Resolution of Universal Speech Attribute Models in Spoken Language Recognition. In *Proc. Interspeech*.

Siniscalchi, S., T. Svendsen, and C. Lee (2008). Toward a detector-based universal phone recognizer. In *Proc. ICASSP*, pp. 4261–4264.

Sproat, R. and O. Fujimura (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of phonetics 21*(291-311).

Stager, C. and J. Werker (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature 388*(6640), 381–2.

Stüker, S., T. Schultz, F. Metze, and A. Waibel (2003). Multilingual articulatory features. *Proc. ICASSP 1*.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. *Proc. ICSLP 2*, 901–904.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics 21*(2), 121–137.

Swadesh, M. (1971). *The origin and diversification of language*. Aldine De Gruyter.

Tajchman, G., D. Jurafsky, and E. Fosler (1995). Learning phonological rule probabilities from speech corpora with exploratory computational phonology. *Proceedings of the 33rd Meeting of the Association for Computational Linguistics, 1995b*.

Tong, R., B. Ma, H. Li, and E. Chng (2009). A target-oriented phonotactic front-end for spoken language recognition. *IEEE Transactions on Audio, Speech, and Language Processing 17*(7), 1335–1347.

Torres-Carrasquillo, P., E. Singer, M. Kohler, R. Greene, D. Reynolds, and J. Deller Jr (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. *Proc. ICSLP*, 89–92.

Tucker, R., M. Carey, and E. Parris (1994). Automatic language identification using sub-word models. *Proc. ICASSP*.

Uemlianin, I. (2005). SpeechCluster: A Speech Data Multitool. *Proc. Lesser Used Languages & Computer Linguistics, European Academy, Bozen/Bolzano, Italy*, 171.

van den Berg, B. (2009). A phonological sketch of Awing. *SIL Cameroon*.

van Niekerk, D. and E. Barnard (2009). Phonetic alignment for speech synthesis in under-resourced languages. *Proc. Interspeech*, 880–883.

Varadarajan, B. and S. Khudanpur (2008). Automatically Learning Speaker-independent Acoustic Subword Units. *Proc. Interspeech*.

Varadarajan, B., S. Khudanpur, and E. Dupoux (2008). Unsupervised learning of acoustic subword units. In *Proc. 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 165–168. Association for Computational Linguistics.

Vu, N. T., F. Kraus, and T. Schultz (2010, Dec). Multilingual a-stabil: A new confidence score for multilingual unsupervised training. In *Spoken Language Technology Workshop*, pp. 183 –188.

Walker, B., B. Lackey, J. Muller, and P. Schone (2003). Language-Reconfigurable Universal Phone Recognition. *Proc. Eurospeech*, 153–156.

Watt, D. and W. Allen (2003). Tyneside English. *Journal of the International Phonetic Association 33*(2), 267–271.

Wells, J. (1982). Accents of English. 3 volumes. *Cambridge University Press*.

Wells, J., W. Barry, M. Grice, A. Fourcin, and D. Gibbon (1992). Standard Computer-Compatible Transcription. *Esprit project 2589 (SAM), Doc. no. SAM-UCL-037*. UCL.

Werner, O. (2000). How to reduce an unwritten language to writing: I. *Field Methods 12*(1), 61.

Wester, M., S. Greenberg, and S. Chang (2001). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. *Proc. Eurospeech*.

Williams, G., M. Terry, and J. Kaye (1998). Phonological Elements As A Basis For Language-Independent ASR. *Proc. ICSLP*.

Williams, L. (1977). The Voicing Contrast in Spanish. *Journal of Phonetics 5*(2), 169–184.

Yang, C. (2009). Nisu dialect geography. *SIL Electronic Survey Reports 7*, 40.

Zavaliagkos, G., J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, and H. Gish (1998). The BBN Byblos 1997 large vocabulary conversational speech recognition system. In *Proc. ICASSP*, Volume 2, pp. 905–908. IEEE.

Zesch, T. and I. Gurevych (2010). Wisdom of crowds versus wisdom of linguists - measuring the semantic relatedness of words. *Journal of Natural Language Engineering. 16*(01), 25–59.

Zissman, M. (1993). Automatic language identification using Gaussian mixture and hidden-Markov models. *Proc. ICASSP 2*.

Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. *Speech and Audio Processing, IEEE Transactions on 4*(1).

Zissman, M. and K. Berkling (2001). Automatic language identification. *Speech Communication 35*(1), 115–124.

Zissman, M. and E. Singer (1994). Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. *Proc. ICASSP 1*.

Appendix A

Additional precision scores

The following tables include all the results for the phonemic analysis experiments, except for phone recognition and alignment. The primary purpose is to include additional precision-based measures such as the average precision.

Algorithm	ROC-AUC	PR-AUC	Avg. precision	Precision
Relative minimal difference (RMD)	80.8%	5.06%		5.48%
Active articulator (AA)	68.8%	2.10%		2.10%
Binary feature edits per phone (BFEPP)	82.1%	4.68%	5.48%	
RMD & AA	81.7%	5.59%		6.06%
Bhattacharyya	57.3%	1.95%	2.26%	
Jeffreys Divergence (JD)	58.9%	2.34%	2.73%	
RMD & JD	81.9%	9.39%	10.44%	
AA & JD	74.3%	3.56%	4.37%	
RMD & AA & JD	82.8%	10.70%	11.82%	
Assimilation criterion	56.7%	2.26%		4.00%
Assimilating features	83.9%	4.00%	4.06%	
Apparent minimal pair (MP)	47.5%	1.21%		1.06%
MP counts	55.9%	1.38%	1.62%	
MP independent counts	53.7%	1.32%	1.55%	
MP counts: all pronunciations	28.3%	0.91%	1.05%	
Chance	50.0%	1.30%		

Table A.1: Evaluation scores on TIMIT for each experiment including precision and average precision

Algorithm	ROC-AUC	PR-AUC	Avg. precision	Precision
Relative minimal difference (RMD)	81.2%	2.68%		2.86%
Active articulator (AA)	74.5%	1.43%		1.43%
Binary feature edits per phone (BFEPP)	87.0%	4.80%	5.31%	
RMD & AA	84.7%	3.91%		4.20%
Jeffreys Divergence (JD)	61.8%	1.06%	1.22%	
RMD & JD	85.6%	5.50%	6.13%	
AA & JD	83.6%	2.98%	3.41%	
RMD & AA & JD	87.7%	8.41%	9.33%	
Assimilation criterion	74.4%	2.23%		2.67%
Assimilating features	77.1%	2.27%	2.12%	
Apparent minimal pair (MP)	64.3%	1.08%		1.11%
MP counts	64.0%	1.07%	1.60%	
MP independent counts	66.0%	1.09%	1.65%	
Chance	50.0%	0.70%		

Table A.2: Evaluation scores on Kua-nsi for each experiment including precision and average precision

Algorithm	ROC-AUC	PR-AUC	Avg. precision	Precision
Relative minimal difference (RMD)	94.7%	16.67%		16.67%
Active articulator (AA)	74.9%	4.04%		4.04%
Binary Feature Edits Per Phone (BFEPP)	99.0%	52.64%	57.91%	
Chance	50.0%	2.10%		

Table A.3: Evaluation scores on French for each experiment including precision and average precision

Appendix B

Phonology sketch of Kua-nsi

This is a draft phonology sketch of Kua-nsi as spoken in the Hedong village, Heqing county, Yunnan province, China. It is based on data from Castro et al. (2010). The analysis is a slight adaptation of Lehonkoski et al. (2010) which is based on the accent spoken in the San'gezhuang village also in Heqing county.

Consonants

pʰ		tʰ	t͡sʰ		t͡ɕʰ			kʰ	
p	f	t	t͡s	s	t͡ɕ	ɕ	ç	k	h
b	v	d	d͡z	z	d͡ʑ	j		g	ɣ
		l						w	
		ʔl̩				ʔj			
m		n			ɲ			ŋ	
ʔm		ʔn			ʔɲ				
		n̩						ŋ̩	

Vowels

i	ɯ	u
i̩	ɯ̩	u̩
	ɤ	o
	ɤ̩	
	a	
	a̩	

Note that [ɯ] in Castro et al. (2010), appears as [ɤ] in Lehonkoski et al. (2010) and vice versa. The vowels combine into a number of diphthongs; the most common are [ia] and [ua].

Syllable structure

The syllable structure is:
V (includes syllabics)
CV
CVV
(CVVV possibly)

There are three phonemic tones for each syllable nucleus; low, mid and high (21, 33, 55)

Allophones

$p^h \rightarrow \widehat{pf^h}$ / *_i̪
p → p͡f / *_i̪
b → b͡v / *_i̪
m → ɱ / *_* (optional) (might be influenced by a following i̪)
h → x / *_[+high, +back] (possibly [x] is underlying)
j → ʑ / *_* (optional) (possibly [ʑ] is underlying, ʑ is infrequent and occurs before front high vowels)

V → [+nasal] / *_* (optional)
V → ʔV / #_, $_
i → ɿ / [+strident,-distributed]_
i̪ → ɿ̪ / [+strident,-distributed]_
ɤ → i̪ / *_* (optional) (transcription by Castro et al. (2010))
or ɯ → i̪ / *_* (optional) (transcription by Lehonkoski et al. (2010))
ɤ̞ → i̪ / *_* (optional) (transcription by Castro et al. (2010))
or ɯ → i̪ / *_* (optional) (transcription by Lehonkoski et al. (2010))
This might be in free variation but the transformation seems more likely to occur when the previous consonant is labial.

Appendix C

IPA mappings for the Brno phone recognisers

The following tables show mappings for the Brno University of Technology phone recognisers (Schwarz et al., 2009) from their own Sampa variant (ASCII) to IPA (Unicode). On the IPA column there is also an indication whether the phone corresponds to a phoneme in the language, or a surface allophone.

BSampa	IPA	BSampa	IPA
pau	[(.)]	c	/ c /
int	[(.)]	J_	/ ɟ /
spk	[(.)]	J	/ ɲ /
p	/ p /	j	/ j /
b	/ b /	k	/ k /
m	/ m /	g	/ ɡ /
F	[ɱ]	N	[ŋ]
f	/ f /	x	/ x /
v	/ v /	h_	/ ɦ /
t	/ t /	i:	/ iː /
t_s	/ t͡s /	i	/ ɪ /
t_S	/ t͡ʃ /	e	/ ɛ /
d	/ d /	e_u	/ ɛ͡u /
d_z	[d͡z]	e:	/ ɛː /
d_Z	[d͡ʒ]	a	/ a /
n	/ n /	a_u	/ a͡u /
r	/ r /	a:	/ aː /
P_	/ r̝ /	u	/ u /
s	/ s /	u:	/ uː /
z	/ z /	o	/ o /
l	/ l /	o_u	/ o͡u /
S	/ ʃ /	o:	/ oː /
Z	/ ʒ /		

Table C.1: Czech symbol mapping to IPA

BSampa	IPA	BSampa	IPA
spk	[(.)]	t1	/ c͡ç /
pau	[(.)]	t1:	/ c͡ç: /
int	[(.)]	d_	/ ɟ͡ʝ /
p	/ p /	d_:	/ ɟ͡ʝː /
b	/ b /	J	/ ɲ /
b:	/ bː /	J:	/ ɲː /
m	/ m /	j	/ j /
m:	/ mː /	j:	/ jː /
F	[ɱ]	k	/ k /
f	/ f /	k:	/ kː /
v	/ v /	g	/ ɡ /
tS	/ t͡ʃ /	N	[ŋ]
tS_	/ t͡ʃː /	x	[x]
t	/ t̪ /	h	/ h /
ts	/ t̪s̪ /	h1	[ɦ]
ts_	/ t̪s̪ː /	i	/ i /
t:	/ t̪ː /	i:	/ iː /
d	/ d̪ /	y	/ y /
dz	/ d̪z̪ /	y:	/ yː /
n	/ n̪ /	e:	/ eː /
n:	/ n̪ː /	_2	/ ø /
r	/ r̪ /	:2	/ øː /
r:	/ r̪ː /	E	/ ɛ /
s	/ s̪ /	A:	/ aː /
s:	/ s̪ː /	u	/ u /
z	/ z̪ /	u:	/ uː /
z:	/ z̪ː /	o	/ o /
l	/ l̪ /	o:	/ oː /
l:	/ l̪ː /	O	/ ɑ /
S	/ ʃ /		
S:	/ ʃː /		
Z	/ ʒ /		

Table C.2: Hungarian symbol mapping to IPA

BSampa	IPA	BSampa	IPA
int	[(.)]	j	/ j /
pau	[(.)]	k	/ k /
spk	[(.)]	k:	[k^j]
p	/ p /	g	/ g /
p:	/ p^j /	g:	[g^j]
b	/ b /	x	/ x /
b:	/ b^j /	x:	[x^j]
m	/ m /	i:	[i]
m:	/ m^j /	i	/ ɪ /
f	/ f /	_1	/ ï /
f:	/ f^j /	e:	[e]
v	/ v /	e	/ ɛ /
v:	/ v^j /	a:	[æ]
t_s	/ $\widehat{ts}$ /	a	/ a /
ts	/ $\widehat{ts}$ /	_1:	[ɨ]
t_S	/ $\widehat{tʃ}$ /	u:	[ʉ]
tS	/ $\widehat{tʃ}$ /	u	/ u /
t	/ t̪ /	o:	/ o /
t:	/ $t̪^j$ /		
d	/ d̪ /		
d:	/ $d̪^j$ /		
n	/ n̪ /		
n:	/ $n̪^j$ /		
r	/ r̪ /		
r:	/ $r̪^j$ /		
s	/ s̪ /		
s:	/ $s̪^j$ /		
z	/ z̪ /		
z:	/ $z̪^j$ /		
l:	/ $l̪^j$ /		
l	/ $l̪^ɣ$ /		
S	/ $ʃ^ɣ$ /		
Z	/ $ʒ^ɣ$ /		
Ss	[ɕː]		

Table C.3: Russian symbol mapping to IPA