**Article:**

Kempton, T. and Moore, R.K. (2013) Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. Speech Communication, 56. 152 - 166. ISSN 0167-6393

https://doi.org/10.1016/j.specom.2013.02.006

# Discovering the phoneme inventory of an unwritten language: a machine-assisted approach

Timothy Kempton, Roger K. Moore

*Department of Computer Science, University of Sheffield, UK*

**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.

It is demonstrated that a machine-assisted approach can make a measurable contribution to a phonemic analysis for all the procedures investigated; phonetic similarity, complementary distribution, and minimal pairs. The evaluation measures introduced in this paper 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.

*Keywords:* phonemic analysis, endangered languages, field linguistics

## 1. Introduction

### *1.1. Motivation*

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).

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).

### *1.2. 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 with example words in Table 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).

[1] 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.

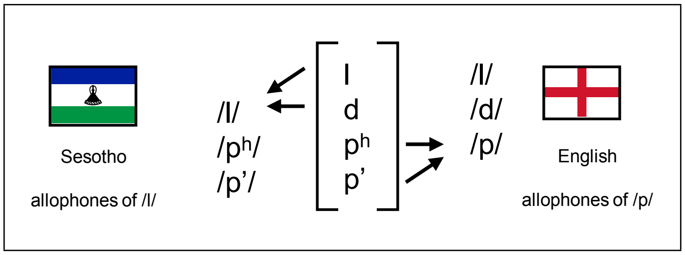

Figure 1: Sesotho and English allophones

| Utterance | Sesotho interpretation | | English (NE) interpretation | |
|---|---|---|---|---|
| [li] | /li/ | them (obj. concord) | /li/ | lea (meadow) |
| [di] | /li/ | them (obj. concord) | /di/ | Dee (UK river) |
| [$\mathrm{p^h}$olo] | /$\mathrm{p^h}$olo/ | ox | /polo/ | polo (sport/mint) |
| [p'olo] | /p'olo/ | Polo (name) | /polo/ | polo (sport/mint) |
| [bolo] | /bolo/ | ball | /bolo/ | Bolo (name) |

Table 1: Sesotho and English perceptions of the same utterance

A phonemic analysis allows an alphabet to be developed for a language, and leads to the important follow-on benefits of literacy and further linguistic analysis. These are discussed further in Kempton (2012). An additional follow-on benefit is the development of speech technology for the language.

### *1.3. Follow-on benefit: 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. 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 received pronunciation (RP) which is traditionally regarded as the prestige English accent. 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 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.

### *1.4. The value of machine-assisted phonemic analysis*

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[2] confirm that such failures can lead to large scale revisions of the phonology; making time estimations difficult. There

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

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. This automated analysis would be particularly valuable when revisions of the analysis are needed, or if the linguist wants to experiment with different hypotheses.

This current paper builds on Peperkamp et al. (2006) which was an investigation into the problem of discovering allophones and their associated rules without knowledge of underlying forms. The previous study was conducted in the context of modelling infant language development. But it is also relevant to a phonemic analysis where the linguist does not know a priori what the underlying forms are. One limitation of this previous study, was that the phonetic data was synthetically derived from a phonemic transcription in the first place. In this current paper the methods of Peperkamp are evaluated on phonemic analysis problems and improved.

### *1.5. 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 2.

#### *1.5.1. The phonetic stage*

One of the first stages in a phonemic analysis is to take an impressionistic phonetic transcription of 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,$p^h$,p'] 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. Aligning the transcription with the waveform is not essential but can be helpful for acoustic analysis such as vowel formant plots (Ladefoged, 2003, p.192).

Automatic phone recognition and alignment has been investigated in Kempton (2012, Ch.4) which led to a tool used for cross-language forced

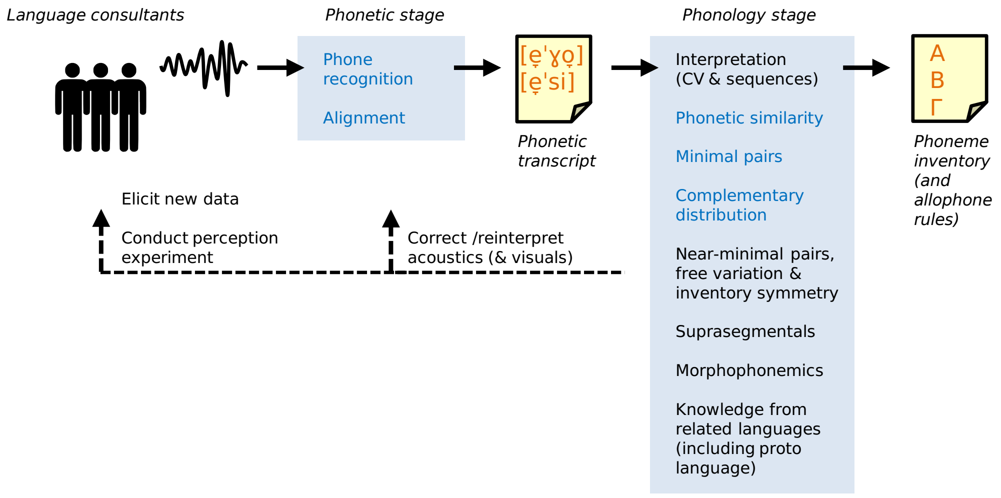
Language consultants
Phonetic stage
Phone recognition
Alignment
[e̞ˈɣo̞]
[e̞ˈsi]
Phonetic transcript
Phonology stage
Interpretation (CV & sequences)
Phonetic similarity
Minimal pairs
Complementary distribution
Near-minimal pairs, free variation & inventory symmetry
Suprasegmentals
Morphophonemics
Knowledge from related languages (including proto language)
A
B
Γ
Phoneme inventory (and allophone rules)
Elicit new data
Conduct perception experiment
Correct /reinterpret acoustics (& visuals)

Figure 2: The stages in a phonemic analysis. Procedures in the phonology stage highlighted in blue (or grey if in monochrome) are those investigated in the paper. Procedures highlighted in the phonetic stage have been investigated in Kempton (2012).

alignment of under-resourced languages (Kurtic et al., 2012).

### *1.5.2. The phonology stage*

Once a detailed phonetic transcript has been attempted, the analysis is primarily phonological.

After deciding on an initial interpretation of ambiguous sounds (see Kempton (2012)), 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. These are shown in Figure 2 and further information can be found in Burquest (2006) and Hayes (2009).

This phonology stage of a phonemic analysis is an iterative one. For example, it is 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 2 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.

### *1.6. 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
- Complementary distribution

- Minimal pairs

To answer the scientific question, a suitable evaluation metric for accuracy is needed; this is introduced in Section 2.3. A search of the literature suggests that these procedures have never been quantitatively evaluated and compared up until now. 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?"*

Following a section on the experimental framework, the subsequent sections are devoted to answering this question by investigating each of the above three procedures. The final section includes a further discussion and a conclusion. The overall aim is not to fully automate the analysis, but to lay the groundwork for a machine-assisted approach that a linguist can use. This is primarily through detecting allophones pairs, a tedious process which would benefit from becoming partly automated.

## 2. Experimental framework

### *2.1. 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 [l$^{ɣ}$] 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. Figure 3 shows some example binary features e.g. [w] has positive labial and dorsal components (lips and tongue body) but a negative coronal component because the tongue tip is not used for this sound. Note that the feature values can be undefined e.g. for [w] the features associated with the tongue blade e.g. lateral are neither + or -. There are 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. This resource is available online[3] and is used in the experiments for this paper. It is also possible to add further features such as tone.

Contour segments also need to be represented. These are sequences of sound that behave phonologically as a single sound such as triphthongs, preglottalized sounds, and tone contours. These can be represented using sequences of binary feature vectors that behave as one unit.

There are many practical advantages to using binary features, but there are also some theoretical shortcomings. One theoretical shortcoming in using binary features is that they are more phonologically motivated than they are phonetically motivated. 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).

### *2.2. 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 paper, 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 sections of this paper.

[3] http://www.linguistics.ucla.edu/people/hayes/IP/features.xls

|  | consonantal | sonorant | continuant | delayed release | approximant | tap | trill | nasal | voice | spread gl | constr gl | LABIAL | round | labiodental | CORONAL | anterior | distributed | strident | lateral | DORSAL | high | low | front | back | tense |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | + | - | - | - | - | - | - | - | - | - | - | + | - | - | - |  |  |  | - | - |  |  |  |  |  |
| b | + | - | - | - | - | - | - | - | + | - | - | + | - | - | - |  |  |  | - | - |  |  |  |  |  |
| m | + | + | - |  | - | - | - | + | + | - | - | + | - | - | - |  |  |  | - | - |  |  |  |  |  |
| f | + | - | + | + | - | - | - | - | - | - | - | + | - | + | - |  |  |  | - | - |  |  |  |  |  |
| v | + | - | + | + | - | - | - | - | + | - | - | + | - | + | - |  |  |  | - | - |  |  |  |  |  |
| θ | + | - | + | + | - | - | - | - | - | - | - | - | - | - | + | + | + | - | - | - |  |  |  |  |  |
| ð | + | - | + | + | - | - | - | - | + | - | - | - | - | - | + | + | + | - | - | - |  |  |  |  |  |
| t | + | - | - | - | - | - | - | - | - | - | - | - | - | - | + | + | - | - | - | - |  |  |  |  |  |
| s | + | - | + | + | - | - | - | - | - | - | - | - | - | - | + | + | - | + | - | - |  |  |  |  |  |
| d | + | - | - | - | - | - | - | - | + | - | - | - | - | - | + | + | - | - | - | - |  |  |  |  |  |
| n | + | + | - |  | - | - | - | + | + | - | - | - | - | - | + | + | - | - | - | - |  |  |  |  |  |
| ɾ | + | + | + |  | + | + | - | - | + | - | - | - | - | - | + | + | - | - | - | - |  |  |  |  |  |
| ɾ̃ | + | + | + |  | + | + | - | + | + | - | - | - | - | - | + | + | - | - | - | - |  |  |  |  |  |
| z | + | - | + | + | - | - | - | - | + | - | - | - | - | - | + | + | - | + | - | - |  |  |  |  |  |
| l | + | + | + |  | + | - | - | - | + | - | - | - | - | - | + | + | - | - | + | - |  |  |  |  |  |
| t͡ʃ | + | - | - | + | - | - | - | - | - | - | - | - | - | - | + | - | + | + | - | - |  |  |  |  |  |
| ʃ | + | - | + | + | - | - | - | - | - | - | - | - | - | - | + | - | + | + | - | - |  |  |  |  |  |
| d͡ʒ | + | - | - | + | - | - | - | - | + | - | - | - | - | - | + | - | + | + | - | - |  |  |  |  |  |
| ʒ | + | - | + | + | - | - | - | - | + | - | - | - | - | - | + | - | + | + | - | - |  |  |  |  |  |
| ɻ | + | + | + |  | + | - | - | - | + | - | - | - | - | - | + | - | - | - | - | - |  |  |  |  |  |
| j | - | + | + |  | + | - | - | - | + | - | - | - | - | - | - |  |  |  | - | + | + | - | + | - | + |
| k | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - |  |  |  | - | + | + | - |  |  |  |
| g | + | - | - | - | - | - | - | - | + | - | - | - | - | - | - |  |  |  | - | + | + | - |  |  |  |
| ŋ | + | + | - |  | - | - | - | + | + | - | - | - | - | - | - |  |  |  | - | + | + | - |  |  |  |
| w | - | + | + |  | + | - | - | - | + | - | - | + | + | - | - |  |  |  | - | + | + | - | - | + | + |
| ʔ | + | - | - | - | - | - | - | - | - | - | + | - | - | - | - |  |  |  | - | - |  |  |  |  |  |
| h | - | - | + | + | - | - | - | - | - | + | - | - | - | - | - |  |  |  | - | - |  |  |  |  |  |
| ɦ | - | - | + | + | - | - | - | - | + | + | - | - | - | - | - |  |  |  | - | - |  |  |  |  |  |

Figure 3: TIMIT consonants with the Hayes (2009) feature set (includes redundancy)

The algorithms are also evaluated on Kua-nsi. Kua-nsi is a Tibeto-Burman language spoken in the Yunnan province of China that currently has no writing system of its own. 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. 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. Audio recordings of the words have been made available to us by the authors.

The experiments in this paper are targeted at the phonology stage of the phonemic analysis process. This means that the input data from the corpora is the manually labelled phonetic transcripts.

The algorithms presented in this paper can process all speech sounds, but the evaluation focuses on consonants. This is because, for the vowel data, there is some variability or uncertainty of vowel ground truth labels in both corpora. 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.

### *2.3. Evaluation measure*

The overall task of detecting allophones can be viewed as an information retrieval problem with allophone pairs representing relevant items and all other phone pairs representing non relevant items. A given algorithm for detecting allophones produces a score so that all the phone pairs scores can be sorted in a ranked list allowing the threshold to be chosen by linguist. Standard information retrieval evaluation tools are then used to measure the performance. 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 4. The ROC-AUC 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 second information retrieval statistic is PR-AUC (precision recall - area under curve). This can be derived by plotting a graph of precision

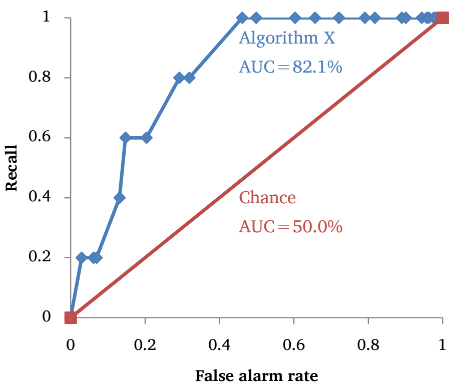

Figure 4: Receiver operating characteristic graph showing area under the curve (ROC-AUC) for an example algorithm and chance.

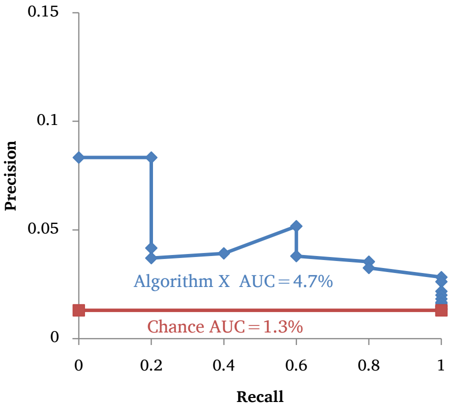

Figure 5: Precision-recall graph showing area under the curve (PR-AUC) for an example algorithm and chance.

against recall and measuring the area under the curve. An example can be seen in Figure 5. Precision ranges from 0 to 1 like recall does but the values are typically much lower for the results in this paper. PR-AUC 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. 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 dataset, 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 would have a PR-AUC value of 0.7%. This is because there is a lower proportion of allophone pairs in the Kua-nsi dataset.

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[4] with AUCCalculator (Davis and Goadrich, 2006).

## 3. Phonetic similarity

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 section different phonetic distance heuristics are evaluated *quantitatively* for their effectiveness in detecting allophones.

### *3.1. Relative minimal difference algorithm*

Peperkamp et al. (2006) makes use of phonetic similarity in an algorithm to model the acquisition 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 4.1. 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

[4] 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 paper. 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.

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

Table 2: Area under the ROC and PR curves for the phonetic similarity algorithms on TIMIT.

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 paper other minimal phonetic distance measures are used so 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 section. 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 section Peperkamp’s phonetic similarity is evaluated for its effectiveness as a standalone process.

The result of the relative minimal difference algorithm as evaluated on TIMIT consonants is shown in the second row of Table 2.

### *3.2. Active articulator algorithm*

A new phonetic similarity detection algorithm is introduced that draws its inspiration from linguists. This is based on the *active articulator* that is used to produce the sound e.g. the lips [labial], the tongue blade [coronal] and the tongue body [dorsal] (Hayes, 2009, p.83). 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 in this case also the 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). The overlap element is included whenever the tongue body [+dorsal] is engaged or the tongue blade is in the palatal region [+coronal, -anterior]. Therefore the active articulator universal set is:

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

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

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

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]:

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

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

**Example 2**, comparing [k] and [ʔ]:

$[\text{k}]_{\text{AA}}$ = {dorsal, dorsal_coronal_overlap}
$[\text{ʔ}]_{\text{AA}} = \emptyset$
The second condition is violated, therefore [k] and [ʔ] are predicted to not necessarily be phonemically distinct.

**Example 3**, comparing [n] and [ŋ]:
$[\text{n}]_{\text{AA}}$ = {coronal, nasal}
$[\text{ŋ}]_{\text{AA}}$ = {dorsal, dorsal_coronal_overlap, nasal}
$[\text{n}]_{\text{AA}} \cap [\text{ŋ}]_{\text{AA}}$ = {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 in Table 2.

Although the results are not particularly high, the active articulator algorithm was found not to miss any allophones but it did have many false alarms. 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.

### *3.3. Binary Feature Edits Per Phone (BFEPP)*

Many different phonetic distance measures have been proposed in the literature Kondrak (2003). Gildea and Jurafsky (1996) created an alignment algorithm and defined the distance between two phones as the number of binary features changed, i.e. the Hamming distance. For example changing [s] to [ʃ] involves changing the two binary features [anterior], and [distributed]; a distance of two.

To handle contour segments, the distance between phone sequences needs to be calculated. Gildea and Jurafsky (1996) use dynamic programming to calculate the cumulative distance for phone sequences where "the cost of insertions and deletions was arbitrarily set at six (roughly one quarter the maximum possible substitution cost)" . In this current study the dynamic programming (with uniform transition penalties) calculates the cumulative

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

Table 3: Area under the ROC and PR curves for the phonetic similarity algorithms on Kua-nsi.

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). As in dialectometry the normalisation can be calculated by dividing by the number of phones in the longest sequence.

The main novelty of this distance algorithm over previous studies is that it is implemented to work with contour segments of any length e.g. triphthongs. This was achieved by running dynamic programming both on the contour segments and the complete phones being compared. It should be noted that the other phonetic similarity algorithms were also extended to handle contour segments. This is described further in Kempton (2012, p.55-59).

The results for BFEPP on TIMIT in Table 2 shows it performs well.

### *3.4. The algorithms applied to Kua-nsi data*

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 Kempton (2012). The results are shown in Table 3. Again the focus was on consonants to make it comparable to previous experiments but this time contour segments such as [ʔ͡n] were included. The different algorithms show the same ranking of performance when compared with the TIMIT results. Again, the active articulator algorithm did not miss any allophones but has many false alarms. The binary feature distance measure is the most successful.

### *3.5. The algorithms applied to a French phone set*

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 paper. However, results from Peperkamp et al. (2006) indicate that the relative minimal difference algorithm has a

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

Table 4: Area under the ROC and PR curves for the phonetic similarity algorithms on the French data.

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 section were evaluated with Peperkamp's French data; specifically 21 consonants plus 9 allophones. Results are shown in Table 4. The results for the three algorithms show an improvement when compared to 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 is probably because the allophones were automatically added to an originally phonemic transcription. 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.

A graph summarising the results of this paper including this section on phonetic similarity, can be found in Figure 6.

## 4. 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.

In this section 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 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

| Algorithm applied to TIMIT | ROC-AUC | PR-AUC |
|---|---|---|
| Jeffreys divergence | 58.9% | 2.3% |
| Assimilation criterion | 56.7% | 2.3% |
| Assimilating features | 83.9% | 4.0% |

Table 5: Area under the ROC and PR curves for the complementary distribution algorithms on TIMIT.

performed on the under-resourced language Kua-nsi. Peperkamp's method to identify the default allophone was also evaluated.

### *4.1. Measuring complementary distribution with Jeffreys divergence*

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.* Similar to Peperkamp et al. (2006), only the following phone is used as the environment for complementary distribution.

### *4.2. Results of using the Jeffreys divergence algorithm on TIMIT*

The results of the Jeffreys divergence algorithm applied to the TIMIT data of 1386 utterances is shown in Table 5. Although these values are shown for the consonants, the analysis has also involved taking account of vowels, utterance boundaries and pauses.

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 should be regarded as the default phone. This technique identified the correct phone for all five allophone pairs within the TIMIT consonant experiment. These five allophone pairs with the default phone appearing first are [t,ɾ] , [d,ɾ], [n,ɾ̃], [t,ʔ], [h,ɦ]. This outcome corresponds to a 3% probability of getting this result by chance (i.e. the probability of choosing the correct phone in the pair five times $= \frac{1}{2^5}$).

| Algorithm applied to Kua-nsi | ROC-AUC | PR-AUC |
|---|---|---|
| Jeffreys divergence | 61.8% | 1.1% |
| Assimilation criterion | 74.4% | 2.2% |
| Assimilating features | 77.1% | 2.3% |

Table 6: Area under the ROC and PR curves for the complementary distribution algorithms on Kua-nsi.

### *4.3. Results of using the Jeffreys divergence algorithm on Kua-nsi*

The experiments performed on TIMIT for complementary distribution were performed on the Kua-nsi data. The results are shown in Table 6. 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.

Relative entropy was used to predict which phone in each phone pair was the default phone. 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.

### *4.4. Using features: two assimilation algorithms*

The Jeffreys divergence algorithm treats all phones as arbitrary symbols and has no knowledge of their features. And yet, features are especially relevant to the sequential constraints imposed on groups of phones in a language e.g. in Kua-nsi (Kempton, 2012, Ch.5). Since phonology rules commonly apply to natural classes (Hayes, 2009, p.71), it is important to integrate features into the algorithms for detecting allophones.

Peperkamp et al. (2006) introduces a method for detecting assimilation i.e. where a segment takes on the characteristics of its phonetic environment. This can help to reveal allophone pairs. For example, in English [n̪] is a dentalized allophone of [n] that occurs before the dental fricative [θ] (Hayes, 2009, p.24). The dental feature is being assimilated from the fricative to the nasal consonant. 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. Going back to the English dentalization example, since the phone pair [n̪θ] frequently occurs together and the dental feature is common to both, the total difference for every single feature summed over the allophone's contexts will be less than the feature difference for [n] and the contexts. This detector does not work well on the TIMIT data using the Hayes feature set (for information on this feature system see Section 2.1) 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 exemplifies one of the limitations of feature modelling and is the reason for the poor result on the second line of Table 5. 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 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 positive but smaller effect on the results of the Kua-nsi data.

Results are shown in Table 5 and Table 6. Overall it can be seen that knowledge of features is beneficial.

### *4.5. Discussion on complementary destribution*

The results in this section show that the application of the Jeffreys divergence algorithm can help detect allophones among the consonants in the TIMIT and Kua-nsi 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).

As in previous work (Peperkamp et al., 2006), on inspecting the data 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.

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.

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 the phonetic/phonological phenomenon underlying complementary distribution.

## 5. 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 section minimal pairs are quantitatively evaluated for their effectiveness in a phonemic analysis. The definition of a minimal pair is “a pair of words differing in only one phoneme” (Clark et al., 2007, p.92). 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.

### *5.1. Existence of putative minimal pairs*

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.

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 2.1) 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. In the experiments reported in this section 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 7 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 sections.

The evaluation measures ROC-AUC and PR-AUC described in Section 2.3 are for evaluating how well an algorithm identifies allophones. However given the probabilistic interpretation of the ROC-AUC measure; the exact

| 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 7: Area under the ROC and PR curves for the minimal pair algorithms on Kua-nsi.

same measure can be given to quantify how well the algorithm identifies phonemically-distinct sounds. 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 paper. The PR-AUC measure however is different and should not be interpreted in the same symmetrical way.

### *5.2. Counts of putative minimal pairs*

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$^{55}$.ɲ̟u$^{33}$] | breast | [ʔu$^{55}$.ɲ̟u$^{33}$] | 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. Also, 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, a count 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.

The results for Kua-nsi are shown in the second row of Table 7. 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.

### *5.3. Using independent counts*

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.

This method of counting independent words, was implemented as a post-process to the Minpair software. The results for Kua-nsi are shown in the bottom row of Table 7. The ROC-AUC measure shows an improvement over those previous results.

### *5.4. Experiments on TIMIT*

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 the rest of this paper, 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. Following the practice of language survey work, the most common pronunciation was chosen for words with multiple pronunciations. For example, 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 are shown in Table 8. 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

| 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 8: Area under the ROC and PR curves for the minimal pair algorithms on TIMIT (the most common pronunciation is used for each word).

| 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] | [ɦə˞d] | heard | [ɦə˞ɾ] | 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 9: Problematic putative minimal pairs in TIMIT that appear to be showing contrast between phones that should be allophones according to the TIMIT documentation

general, performing little better than chance i.e. the ROC-AUC value is near to 50% (see Section 2.3).

Investigating the data revealed that the poor result was due to a number of known allophones that had putative minimal pairs. The putative minimal pairs for sounds described as allophones in the TIMIT documentation (Garofolo et al., 1993) are listed in Table 9.

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 tap 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 due 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 word 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 9, 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. This is in line with the observation that "the discovery of phonetically minimal pairs does not necessarily permit an immediate conclusion about underlying phonological contrast" (Postal, 1968, p.28). 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. Discussion and Conclusion

### *6.1. Summary of results*

Figure 6 summarises the results for the three procedures; phonetic similarity, complementary distribution and minimal pairs. As discussed earlier the ROC-AUC values should be regarded as the primary evaluation measure for comparing algorithms. All the different procedures show a better than chance performance except for the putative minimal pair algorithm when applied to the TIMIT data.

The phonetic similarity algorithms investigated in Section 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 (see Section 3.5).

The complementary distribution algorithms investigated in Section 4 for TIMIT and Kua-nsi are shown in the centre of the bar charts in Figure 6. 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 6 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 4.2 and 4.3).

The minimal pair algorithms investigated in Section 5 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

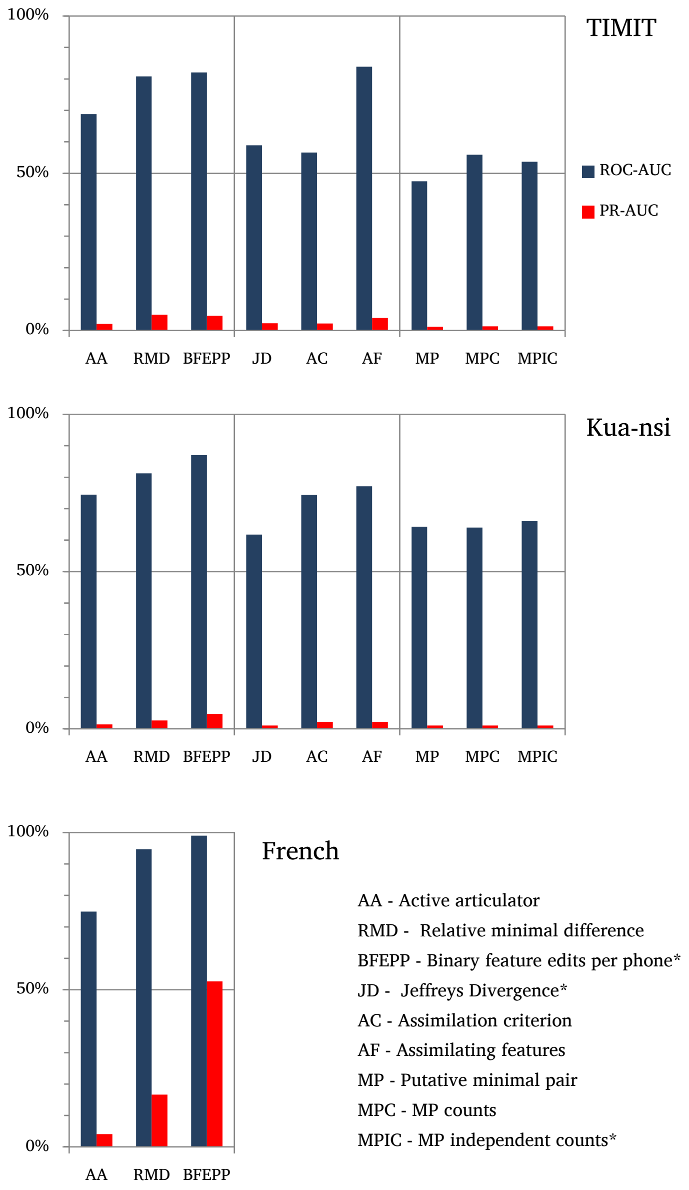

Figure 6: Graph showing summary results of phonemic analysis procedures. The vertical lines indicate the groupings of algorithms into the phonetic similarity, complementary distribution and minimal pair procedures. The horizontal line at 50% indicates the chance level for the ROC-AUC values (the chance values for PR-AUC are not shown). *Algorithms with an asterisk are those that best represent each procedure.

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.

### *6.2. 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 6 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 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, Jeffreys 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.

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; it is machine-assisted phonemic analysis not machine-automated phonemic analysis.

As explained in Section 5.1, the ROC-AUC statistic used in this paper 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[5] is not specifically defined. However the findings in this paper do cast doubt on any premise that minimal pairs alone are the most effective method for detecting phonemically distinct phones in a phonemic analysis.

## 7. Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Research Council (EPSRC grant number EP/P502748/1). We are grateful to Andy Castro for providing the Kua-nsi recordings, Mary Pearce and Cathy Bartram for helpful discussions about phonemic analysis heuristics, and Sharon Peperkamp for answering our questions regarding previous experiments. We would also like to thank Nic de Vries for comments on the application of the tools. Image component credits: Zscout370 (Lesotho flag, CC BY-SA 2.5).

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

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

[5] "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)

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

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

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

Crystal, D., 2000. Language death. Cambridge Univ Press.

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

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

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

Fitt, S., Isard, S., 1999. Synthesis of regional English using a keyword lexicon, in: Sixth European Conference on Speech Communication and Technology.

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

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

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

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

Hayes, B., 2009. Introductory phonology. Wiley-Blackwell.

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

Hockett, C., 1955. How to learn Martian. Astounding Science Fiction , 97–102.

Huckvale, M., 2004. ACCDIST: a metric for comparing speakers' accents, in: Proc. ICSLP.

Jeffreys, H., 1948. Theory of probability. Oxford University Press. 2 edition.

Kempton, T., 2012. Machine-assisted phonemic analysis. Ph.D. thesis. University of Sheffield.

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

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

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

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

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

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

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

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.

SIL, 2008. Phonology Assistant v3.0.1. SIL International [Software].

Siniscalchi, S., Svendsen, T., Lee, C., 2008. Toward a detector-based universal phone recognizer, in: Proc. ICASSP, pp. 4261–4264.

Sproat, R., Fujimura, O., 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. Journal of phonetics 21.

Wells, J., 1982. Accents of English. 3 volumes. Cambridge University Press .

Williams, L., 1977. The Voicing Contrast in Spanish. Journal of Phonetics 5, 169–184.