Cross-language phone recognition when the target language phoneme inventory is not known

Timothy Kempton, Roger K. Moore, Thomas Hain

Department of Computer Science, University of Sheffield, UK

{t.kempton,r.k.moore,t.hain}@dcs.shef.ac.uk

Abstract

Cross-language speech 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. In the work reported here, phone recognisers are evaluated on a cross-language task with minimum target knowledge. A phonetic distance measure is introduced for the evaluation, allowing a distance to be calculated between any utterance of any language. This has a number of spin-off applications such as allophone detection, a phone-based ROVER approach to recognition, and cross-language forced alignment. Results show that some of these novel approaches will be of immediate use in characterising languages where there is little phonological knowledge.

Index Terms: universal phone recognition, cross-language speech recognition, forced alignment, under-resourced languages

1. Introduction

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 [1]. 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 [2], [3]. However, for many languages, particularly languages that do not have writing systems, the phoneme inventory is not known. This is the case for hundreds, if not thousands, of languages since only 42% of languages are known to have writing systems [4]. Deriving an inventory of phonemes (contrastive sounds) from phones (sounds with an unspecified contrastive status) is a non-trivial task [5], [6]. As far as the authors are aware, there has only been one paper that has explicitly addressed the problem of cross-language phone recognition with minimal target knowledge. Walker et al. [7], 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 sounds to distinguish. Ideally every possible phonetic contrast that might occur in a language needs to be detected.

Despite the difficulties, there are potentially great rewards to be gained by addressing this problem of simulating an expert phonetician. A truly universal phone recogniser for fine phonetic detail could assist in the process of phonemic analysis, characterising the phoneme inventory and the associated models of allophonic variation [6]. This could facilitate the development of a writing system for a new language, further linguistic analysis, and customised speech technology. For languages, dialects and accents that possess a writing system, such an analysis could lead to better models of pronunciation variation.

In this paper, state-of-the-art phone recognisers developed by Schwarz et al. [8], for Czech, Hungarian and Russian, were evaluated on a target language where the phoneme inventory is not assumed to be known. This type of evaluation is characterised by the ground truth containing fine phonetic detail. To improve the evaluation, a new measure in addition to the Phone Error Rate (PER) was used. This is the Binary Feature Errors Per Phone (BFEPP) measure which represents a phonetic distance between utterances of any language. This measure led to further novel developments such as improved allophone detection, a ROVER [9] approach that is phone-based rather than word-based, and cross-language forced alignment. Most of the resources and software used in this paper are publicly available.

2. Phonetic distance and alignment

2.1. 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. Initial documentation of the language has been completed by Castro et al. [10]. 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 Kua-nsi corpus is ideal for evaluating cross-language phone recognition with minimum target knowledge. However, before a full scale evaluation, the audio needs to be trimmed so that it matches the transcriptions. 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.

A recording of the first word in the list, the Kua-nsi word for *sky* in the Hedong dialect, is 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 l̍ a m u (.)]

The above IPA notation includes the pause symbol from the extended IPA [11] and the superscript numbers follow the Chinese phonetic convention of labelling 1 as a low tone and 5 as a high tone. Phone recogniser labels were converted from ASCII to IPA Unicode, using the SAMPA specification [12] and the documented phonology of the relevant language [11], [13].

2.2. Phone Error Rate (PER)

The standard tool to calculate word error rate in speech recognition, SCLITE, [14] can be used to calculate the Phone Error Rate (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. A very similar measure is used in dialectometry; in this case the distance is usually normalised by the length of the longest phone sequence [15], [10].

The use of dynamic programming to align two sequences was later extended to multiple sequences for combining different speech recognisers. This is called ROVER [9] and is explored further in Section 3.3.

When calculating PER, only exact matches are allowed, e.g. in the above Czech example there are six differences when compared to the nine phones in the reference transcription (ignoring the pauses) giving a 67% PER. It is not apparent from such a high error rate how close the substituted phones were to the reference transcription. Also not all applications of cross-language phone recognition require an exact match. It was considered to be more informative to use an error rate that takes account of phonetic distance.

2.3. Binary Feature Errors Per Phone (BFEPP)

Many different phonetic distance measures have been proposed in the literature [15]. Gildea and Jurafsky [16] 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.

One issue with 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 [17]. Even though these sounds have the same voicing, a direct comparison of the symbols suggests a difference of one binary feature; [voice]. This is partly due to the limited detail inherent in phonetic transcripts and this example suffers a greater penalisation with the PER measure. One alternative distance measure which is more phonetically motivated is ALINE [15]. However this was not designed to solve the above problem and appears to require parameter tuning for the task involved.

In the study reported here it was decided that the binary feature approach of Gildea and Jurafsky would be adapted. The main appeal of binary features is their simplicity in implementation, and their flexibility in representing speech sounds with multiple articulations such as apical vowels. 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 [3].

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)" [16]. In this current study 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).

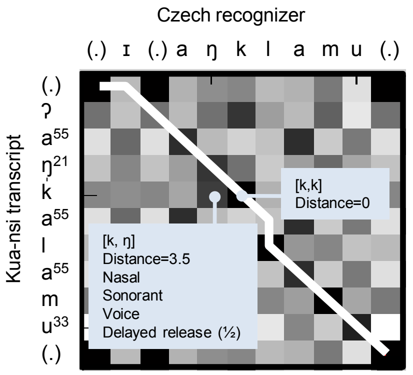

Figure 1: *The distance between two phones is the number of binary feature edits. The cumulative phonetic distance between two phone sequences is calculated with dynamic programming.*

In dialectometry the normalisation can be calculated by dividing by the number of phones in the longest sequence. When measuring the errors of a phone recogniser, 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 2.1, the BFEPP measure gives a value of 3.3 A similar example illustrates the dynamic programming in Figure 1. The first vowel picked up by the Czech recogniser was the interviewer saying a word in Chinese.

The 28 binary features used in this study are defined by Hayes [5]. This was extended slightly to include four binary features for tone. Optionally, a percentage score can be given by dividing BFEPP by the number of features in the binary feature system. Underspecified features are assigned a value halfway between the corresponding +/- values. The main novelty of this distance algorithm is that it is implemented to allow contour segments of any length. This allows the use of phones such as triphthongs, preglottalized sounds, and tone contours. This was achieved by representing phones with multiple binary feature vectors and then using dynamic programming for individual phone comparisons as well as the phone sequence comparisons.

2.4. Allophone detection

The BFEPP distance is effective for detecting allophones; sounds in languages that do not function contrastively. In previous experiments [6] the best performance of a single algorithm gave an accuracy of 80.8% ROC-AUC (area under the ROC curve). The simpler BFEPP measure outperforms this at 82.1% ROC-AUC (with an average precision at 0.055).

3. Cross-language experiments

3.1. 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 current evaluation. The TIMIT corpus [18] was chosen because it is one of the only corpora that contain a large number of manually annotated allophones. The evaluation was

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 1: *Cross-language phone recognition on TIMIT*

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 [8] 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 Czech, Hungarian, and Russian phone recognisers process telephone bandwidth audio, so the TIMIT test data was downsampled.

3.2. Direct cross-language phone recognition

Cross-language phone recognition was first performed directly. Results are shown in Table 1. The Czech and Hungarian recognisers show a greater accuracy than the 87% PER equivalent in Walker et al. [7] 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.

3.3. Phone-based ROVER

ROVER (Recogniser Output Voting Error Reduction) [9] is an algorithm that combines 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.

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.

The phone-based ROVER system was implemented using the SRILM [19] toolkit's nbest-lattice program. Phonetic alignment, similar to that described in Section 2, was performed through a specially created dictionary file to give Hamming distances between phones. This required a corresponding modification to the insertion and deletion penalty in the software.

The ROVER results are shown in Table 1. The simple vote recogniser, is very similar to the Nist1 vote in the original ROVER study [9]. 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).

The results appear disappointing, with none of the scores improving on the best component 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. 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 they are measuring different types of errors.

Visual inspection suggested that phonetic alignment did improve results, but that all methods produced many alignment errors. An attempt was made to tune the insertion deletion penalty, and although this did reduce the alignment errors by approximately half, there was only a small improvement in recognition rates and they still did not surpass the Czech performance.

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

Table 2: *Cross-language forced alignment on TIMIT*

3.4. Cross-language forced alignment

The analysis of under-resourced languages often requires the alignment of phonetic transcripts with audio e.g. for acoustic analysis. This is usually achieved with forced alignment. 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.

To address this problem, the concept of cross-language forced alignment is introduced. This is similar to cross-language phone recognition except a 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 performance of each recogniser was evaluated with software provided by Hosom [20]. This gives the standard forced alignment evaluation error; the proportion of boundaries placed more than 20ms away from the true boundary, and is shown in Table 2. The performance ranking for the three languages is similar to the recognition task.

Two combination methods were investigated, taking the mean of the boundaries, and taking the median of the boundaries. Table 2 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 with the Greenhouse-Geisser correction. The factor *recogniser* had 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 Vs mean combination (p>0.99) and Hungarian Vs 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 has the potential to half the time needed i.e. errors can be quickly identified visually and highlighted by low confidence scores.

3.5. Future work

If the success of combining recognizers for forced alignment is to be followed by a similar ROVER success, both alignments and combination methods will need to be improved.

There are many options for combining phones from different recognisers. 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 2.1 was processed through the different recognisers and the ROVER 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 2.1; 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. Conclusions

This paper has introduced the BFEPP measure for calculating the distance between phone sequences of any language. This measure gives alignments and distances that are both accurate and intuitive. There are a number of spin-off applications from this development including allophone detection, the phone-based ROVER combination method, and cross-language forced alignment.

Using the BFEPP measure for allophone detection has outperformed more complex techniques. The evaluation of cross-language phone recognition using the phone recognisers developed by Schwarz et al. [8] showed a result that compared favourably with a previous study where the phoneme inventory was not known. Unlike word-based ROVER we were unable 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 are needed. Improving the alignments and using different combination methods is a priority for future work.

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. The accuracy level is high enough to be useful to linguists analysing languages where there is little phonological knowledge. Combining different phone recognisers is shown to improve performance.

5. 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, Matthew Kempton for the statistical analysis and John-Paul Hosom, Andreas Stolcke & Petr Schwarz for assistance in using their software.

6. References

[1] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Academic Press, 2006.

[2] T. Schultz and A. Waibel, "Multilingual and Crosslingual Speech Recognition," *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, pp. 259–262, 1998.

[3] S. Siniscalchi, T. Svendsen, and C. Lee, "Toward a detector-based universal phone recognizer," in *Proc. ICASSP*, 2008, pp. 4261–4264.

[4] M. P. Lewis, *Ethnologue: Languages of the world*. SIL International, 2009; Personal communication with the editorial team, 2010.

[5] B. Hayes, *Introductory phonology*. Wiley-Blackwell, 2009.

[6] T. Kempton and R. K. Moore, "Finding allophones: an evaluation on consonants in the TIMIT corpus," *Proc. Interspeech*, pp. 1651–1654, 2009.

[7] B. Walker, Lackey, B. Muller, and P. Schone, "Language-Reconfigurable Universal Phone Recognition," *Proc. Eurospeech*, pp. 153–156, 2003.

[8] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, 2009.

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

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

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

[12] J. C. Wells *et al.* SAMPA - computer readable phonetic alphabet, 2003.

[13] I. Maddieson, *Patterns of Sound*. Cambridge, 1984.

[14] NIST, "Sclite v2.4 score speech recognition system output," 2009.

[15] G. Kondrak, "Phonetic alignment and similarity," *Computers and the Humanities*, vol. 37, no. 3, pp. 273–291, 2003.

[16] D. Gildea and D. Jurafsky, "Learning Bias and Phonological-Rule Induction," *Computational Linguistics*, vol. 22, no. 4, pp. 497–530, 1996.

[17] L. Williams, "The Voicing Contrast in Spanish." *Journal of Phonetics*, vol. 5, no. 2, pp. 169–184, 1977.

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

[19] A. Stolcke, "SRILM-an extensible language modeling toolkit," *Proc. ICSLP*, vol. 2, pp. 901–904, 2002.

[20] J. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, no. 4, pp. 352–368, 2009.