An acoustic distance measure for automatic cross-language phoneme mapping

Jayren J Sooful, Elizabeth C Botha

Department of Electrical, Electronic and Computer Engineering
University of Pretoria, 0002, South Africa.
liesbeth.botha@eng.up.ac.za

Abstract

This paper explores an automated approach to mapping one phoneme set to another, based on the acoustic distances of the individual phonemes. The main goal of this investigation is to automate the technique for creating initial/baseline acoustic models for a new language. Using this technique, it would be possible to rapidly build speech recognition systems for a variety of languages. A subsidiary objective of this investigation is to compare different acoustic distance measures and to assess their ability to quantify the acoustic similarity between phonemes. The distance measures that were considered for this investigation are the Kullback-Leibler measure, the Bhattacharyya distance metric, the Mahalanobis measure, the Euclidean measure, the L2 metric and the Jeffreys-Matusita distance. Both the TIMIT and SUN Speech corpora were used. It was found that by selecting an appropriate distance measure, an automated procedure to map phonemes from a source language (English) to a target language (Afrikaans) can be applied, with recognition results comparable to a manual mapping process undertaken by a phonetic expert.

1. Introduction

This paper presents a technique for building the initial acoustic phoneme models of a hidden Markov model (HMM) in a new/target language (Afrikaans) using acoustic models trained in a source language (English). It is accomplished by finding an appropriate acoustic distance measure in order to automatically map the phoneme set of the source language to the phoneme set of the new language. This technique is especially relevant for tasks where an automatic speech recognition system has already been trained in the source language. Very often, much less training data for the new language is available for building a completely new recogniser.

When a recognition system is developed for a new language (either exclusively for the new language or for the new language in addition to existing languages) the recognition system optimised for the source language has to be adapted to the characteristics of the new language. These techniques are relevant to any new language that has a high degree of overlap with the source language in terms of phonemes.

The experiments are carried out using the TIMIT English database [13] and the Afrikaans segment of the SUN Speech English-Afrikaans corpus [14].

Different acoustic measures are used to compute the acoustic similarity between the TIMIT phoneme models and the SUN Speech phoneme models. The automated approach is then compared to a manual phoneme-mapping procedure carried out by a phonetic expert [3]. The trained TIMIT-based English recogniser is used as the basis for this comparison.

The organisation of this paper is as follows. Section 2 discusses the distance measures that are used in the investigation. The experiments themselves are described in Section 3. The results for the experiments are presented in Section 4 while the conclusions are discussed in Section 5.

2. Distance Measures

A variety of distance-based algorithms exist to compute the distances between Gaussian distributions obtained for each phoneme model.

Let μ_i and Σ_i represent the feature mean vector and covariance matrix respectively for a Gaussian distribution $\boldsymbol{i}$.

A popular distance metric that has been used previously in calculating the distance between two models is the Kullback-Leibler measure [1], which is given by:

$$D_{KL} = \frac{1}{2}(\mu_2 - \mu_1)^T \left[\Sigma_1^{-1} + \Sigma_2^{-1}\right](\mu_2 - \mu_1) + \frac{1}{2} tr\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right) \qquad (1)$$

The Bhattacharyya distance metric [2, 3] has been extensively used to obtain the distance between phoneme models of different languages. This distance measure is given by:

$$D_{Bha} = \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1}(\mu_2 - \mu_1) + \frac{1}{2}\log\frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad (2)$$

The first term gives the class separability as a result of the class means, while the second term gives the class separability between the class covariance matrices.

The Mahalanobis distance metric has also been used as a distance classifier. It has the advantage that by utilising the information available in the covariance matrices, it takes the variability between the models to be compared into account. The Mahalanobis distance [1] is given by the equation:

$$D_{Mah} = \frac{1}{n}\left(\mu_2 - \mu_1\right)^T \left(\Sigma_1 \, \Sigma_2\right)^{-1} \left(\mu_2 - \mu_1\right) \tag{3}$$

The one-dimensional Euclidean measure has also been used to calculate inter-class distances [1, 5]. This geometric measure is given by:

$$D_{Euc} = \left(\mu_2 - \mu_1\right)^T \left(\mu_2 - \mu_1\right) \tag{4}$$

Another popular measure is the L2 distance [1]. A closed form of the L2 distance measure exists if Gaussian distributions are assumed. The L2 distance measure then reduces to:

$$\begin{aligned}(D_{L2})^2 = &\prod_{k=1}^{n} \frac{1}{2\sigma_{1k}\sqrt{\pi}} + \prod_{k=1}^{n} \frac{1}{2\sigma_{2k}\sqrt{\pi}} + \\ &2\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{1k}^2 + \sigma_{2k}^2}} \exp\left[\frac{1}{2}\frac{\left(\mu_{1k}\frac{\sigma_{2k}}{\sigma_{1k}} + \mu_{2k}\frac{\sigma_{1k}}{\sigma_{2k}}\right)^2}{\sigma_{1k}^2 + \sigma_{2k}^2} - \left(\frac{\mu_{1k}^2}{\sigma_{1k}^2} + \frac{\mu_{2k}^2}{\sigma_{2k}^2}\right)\right]\end{aligned} \tag{5}$$

The final distance measure that was used during this investigation is the Jeffreys-Matusita distance measure [12], which is closely related to the Bhattacharyya distance. It reduces to the following expression if a Gaussian distribution is used:

$$D_{JM} = \sqrt{2\left(1 - e^{-\alpha}\right)} \tag{6}$$

where α is given by the value of the Bhattacharyya distance in Equation (2).

3. Experiments

The experiments are carried out using the TIMIT English database [13] and the SUN Speech English-Afrikaans corpus [14]. Only the SI (phonetically-diverse) and SX (phonetically-compact) TIMIT sentence sets were used. The TIMIT database contains about 80% more speech data than the English part of the SUN Speech database. There are 39 different phonemes listed in the TIMIT database (including the silence model) and a total of 59 phonemes used in the labelling of the SUN Speech database.

For the purposes of these experiments the [cl] silence model in TIMIT was mapped directly to the [sil] model in the SUN Speech database. Moreover, 6 phoneme classes are found only in the English segment of the SUN Speech corpus, not in Afrikaans. Since only the Afrikaans data was used, the phoneme-mapping experiments involve 38 TIMIT "base" or "reference" phonemes and 52 SUN Speech phonemes.

The Hidden Markov Toolkit (HTK) version 3 was used to conduct all the experiments [11]. Standard left-to-right Continuous Hidden Markov Models were used. Only diagonal covariance matrices were used. The HTK configuration file was set up to calculate 12 Mel-Frequency Cepstral Coefficients (MFCCs), a log energy measure, delta coefficients and delta-delta coefficients as well. Cepstral mean normalisation [11] was also performed to compensate for audio effects. This is especially relevant in this set of experiments where two independent speech databases are used.

Different acoustic measures are used to compute the acoustic similarity between the TIMIT phoneme models and the SUN Speech phoneme models. The automated approach is then compared to a manual phoneme-mapping procedure carried out by a phonetic expert [3]. The six different distance measures described in Section 2 are used. Since the Gaussian models are multi-mixture (up to four mixtures per state were used in the experiments) and multi-state (three states), the distance between two phones was calculated per mixture per state and added for a total distance measure.

For the purposes of the experiment, only the Afrikaans part of the SUN Speech database was utilised. This was done to mimic practical instances where a small amount of data is available for the new language (Afrikaans) and where a fully trained recogniser already exists for a base language (English).

The SUN Speech database consists of two phonetically rich sentence sets (693 sentences in total) spoken by male and female speakers. The database consists of speakers who spoke both sets and either one of the two sets. Table 1 summarises this breakdown.

Table 1: Details of the SUN Speech Afrikaans database

Sentence sets spoken	Number of sentences spoken by males	Number of sentences spoken by females	Total
1	194	49	243
2	140	10	150
1 & 2	80	220	300
			693

In order to have a representative amount of data for training and testing purposes, the data was split into a 70% training-30% test ratio, maintaining the split based on the information in Table 1 as well. Any speaker who spoke both sentence sets will be found exclusively in either the training or test sets, not in both.

Table 2 below describes the SUN Speech training and test sets used.

Table 2: SUN Speech training and test data

	Training sent.	Test sent.
Male speakers Set 1	134	60
Male speakers Set 2	100	40
Male speakers Set 1 & 2	60	20
Female speakers Set 1	40	9
Female speakers Set 2	10	0
Female speakers Set 1 & 2	140	80
TOTAL	484 (69.8%)	209(30.2%)

In the experiments that were conducted, the following performance criteria are used:

$$\%\ \textit{Correct labels} = \frac{\textit{number of correct labels}}{\textit{total number of labels}}\ \textit{X 100\%}$$

$$\%\ \textit{Accuracy} = \frac{\textit{number of correct labels - insertions}}{\textit{total number of labels}}\ \textit{X 100\%}$$

4. Results

The English recogniser was trained using the TIMIT SI and SX training data and tested using the TIMIT test data. The number of mixtures per state was incremented (in steps of one) from one to four.

The Afrikaans phoneme recogniser was trained using 70% of the available SUN Speech Afrikaans data. Here again, the number of mixtures used was increased from one to four. As a benchmark, the Afrikaans recogniser was tested with the remaining 30% Afrikaans data. The recogniser correctly identified 67.35% of the Afrikaans phonemes with an accuracy of 62.70%.

Each of the six distance measures described in Section 2 was calculated for every TIMIT-SUN Speech phoneme pair (a 38-by-52 distance matrix was computed for each distance measure). Each SUN Speech phoneme was then mapped to the closest TIMIT phoneme (no distance threshold was applied).

The mapped SUN Speech Afrikaans test data was then recognised by the trained TIMIT-based English recogniser. The percentage of correctly recognised phonemes appear in Table 3.

Table 3: Performance of TIMIT-based recogniser on Afrikaans data per distance measure

Dis. Mea.	Man. (Exp.)	KL	Bhat.	Mah.	Euc.	L2	JM
% Corr	32.4	27.1	27.3	13.8	29.5	11.0	32.7
% Acc.	18.2	13.7	13.0	2.2	14.8	0.9	19.1

The results shown in Table 3 are not comparable to the 67.35% correctly identified phonemes obtained when the Afrikaans recogniser was tested with the remaining 30% Afrikaans test data. This 30% test data is a subset of the SUN Speech database, and is thus very similar to the training data. Moreover, the approach of training a new language recogniser from scratch with limited amounts of speech data available is not practical in a continuous speech recognition system. The purpose of this investigation was only to find the optimal distance measure for mapping phonemes.

The techniques listed in [10] and [3] will have to be used to utilise the available new language (Afrikaans) data optimally. These techniques include:

- pooling multilingual data to construct multilingual phone models
- adapting the models trained on the base language (English) using the target language data (Afrikaans)
- training models on multilingual pooled data, and then adapting them using the target language data
- data augmentation by transforming the base language data to better match the target language data, pooling this transformed data with target language data, and then performing further adaptation using the target language.

From Table 3, it can be seen that all the distance metrics used, barring the Mahalanobis and L2 measures, had comparative performance to the manual mapping performed by the phonetic expert. In fact, the Jeffreys-Matusita measure actually outperformed the results achieved by the manual mapping process.

Tables 4 and 5 below lists the two phoneme classes with the best and worst correct phoneme recognition percentages per distance measure.

Table 4: Best recognition performance listed as a percentage of correctly recognised phonemes per distance measure

Dis. Mea.	Man. (Exp.)	KL	Bhat.	Mah.	Euc.	L2	JM
Best	f 81.0	sh 89.7	sh 86.8	cl 77.2	cl 75.8	cl 76.7	sh 89.5
2nd Best	ay 76.1	f 81.3	f 79.3	ih 60.1	z 70.9	aw 44.4	cl 78.9

From Table 4 it is evident that the [f], [sh] and [cl] (silence) models were recognised the best. In general, the fricative sounds were recognised the best by the TIMIT-based recogniser.

Table 5: Worst recognition performance listed as a percentage of correctly recognised phonemes per distance measure

Dis. Mea.	Man. (Exp.)	KL	Bhat.	Mah.	Euc.	L2	JM
Worst	r 5.3	aw 5.7	aw 5.3	ch 1.8	ch 3.0	uh 0	aw 3.4
2nd Worst	t 8.1	p 5.9	p 12.9	jh 2.1	t 6.0	uw 0	p 8.8

Generally, the [aw], [p], [t] and [ch] models displayed the poorest recognition results. Overall, the “stop” class of phoneme models tended to have the worst recognition results.

According to the work done in [3], phonetically there are just two Afrikaans phoneme classes in the SUN Speech database that do not appear in the English part of the database (these are represented by the [R] and [r] phonemes or by their numerical ASCII codes of 82 and 94 respectively). These were grouped into a single [r] class during the manual mapping procedure. It should be noted that the recognition results for this phoneme model were the poorest, indicating that the manual mapping for these two phoneme classes is not a true indication of their acoustic nature.

5. Conclusions

This investigation has shown that an automatic phoneme mapping procedure can be used to map phonemes from a new target language to a base language for which a trained recogniser already exists.

These experiments have also demonstrated that the choice of acoustic distance measure does influence the results obtained. Four out of the six distance measures compared favourably with the manually undertaken phoneme mapping of the phonetic expert.

The approach followed here can be extended to map between phonemes where same-language speech databases do not follow a consistent phoneme labelling schema.

The investigation methodology could be improved by the addition of a threshold condition that compares two phoneme models and maps the phoneme models to each other only if the distance between them is below a predefined threshold.

Although the performance of the English recogniser on Afrikaans data does not compare to it's recognition performance on English data, it should be borne in mind that the English-based recogniser performed reasonably well when presented with a new language, without any data pooling, model-adaptation or retraining.

6. Acknowledgements

The authors would like to thank Derik Thirion for his assistance regarding the workings of the HTK Toolkit, and Darryl Purnell for his help in converting the SUN Speech raw audio files. The authors would also like to thank the University of Stellenbosch for the use of their SUN Speech database.

7. References

[1] L. Couvreur and J. Boite "Speaker Tracking in Broadcast Audio Material in the Framework of the THISL Project," *Proceedings of 1999 Workshop on Accessing Information in Spoken Audio* (ESCA-ETRW), pp.84-89, April 1999..

[2] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," *Proc. ICSLP '96*, Vol. 4, (Philadelphia, PA), pp. 2005-2008, Oct. 1996.

[3] C. Nieuwoudt, "Cross-language Acoustic Adaptation for Automatic Speech Recognition," *PhD Thesis*, University of Pretoria, South Africa, April 2000.

[4] M. Falkhausen, H. Reininger and D. Wolf "Calculation of Distance Measures between Hidden Markov models," *ESCA-NATO Workshop on Multilingual Interoperability in Speech Technology*, Leusden, The Netherlands, Sept. 1999.

[5] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[6] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. COM-15, pp. 52–60, 1967.

[7] J. Kohler "Comparing three methods to create Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks," *ESCA-NATO Workshop on Multilingual Interoperability in Speech Technology*, Leusden, The Netherlands, pp. 79-84, 1999.

[8] J.W.F. Thirion "Phoneme Recognition with HTK on the TIMIT Database," *African Speech Technology Technical Report*, University of Pretoria, South Africa, 23 Feb. 2001.

[9] J.W.F. Thirion "HTK Installation Instructions under Windows NT," *African Speech Technology Technical Report*, University of Pretoria, South Africa, 23 Feb. 2001.

[10] C. Nieuwoudt and E. Botha, "Cross-language use of acoustic information for automatic speech recognition," Accepted for publication in *Speech Communication*, Aug. 2000.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book, " HTK Version 3.0, Microsoft Corporation, July 2000.

[12] S. Pissarra, C. Ribeiro, L.V. Dutra, C.D. Renno and J.V. Soares, "Culture classification using polarimetric information from SIR-C/X-SAR mission : Bebedouro region, Brazil," Technical Report SIR-C/X-SAR, 1996.

[13] ARPA, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," NIST Speech Disc CD1-1.1, Dec. 1990.

[14] Department of Electrical and Electronic Engineering - University of Stellenbosch, "The SUN Speech Database," 1997.