Fusion of Contrastive Acoustic Models for Parallel Phonotactic Spoken Language Identification

Khe Chai Sim and Haizhou Li

Institute for Infocomm Research, Singapore
{kcsim,hli}@i2r.a-star.edu.sg

10.21437/Interspeech.2007-72

Abstract

This paper investigates combining contrastive acoustic models for parallel phonotactic language identification systems. PRLM, a typical phonotactic system, uses a phone recogniser to extract phonotactic information from the speech data. Combining multiple PRLM systems together forms a Parallel PRLM (PPRLM) system. A standard PPRLM system utilises multiple phone recognisers trained on different languages and phone sets to provide diversification. In this paper, a new approach for PPRLM is proposed where phone recognisers with different acoustic models are used for the parallel systems. The STC and SPAM precision matrix modelling schemes as well as the MMI training criterion are used to produce contrastive acoustic models. Preliminary experimental results are reported on the NIST language recognition evaluation sets. With only two training corpora, a 12-way PPRLM system, using different acoustic modelling schemes, outperformed the standard 2-way PPRLM system by 2.0-5.0% absolute EER.

Index Terms: language identification, precision matrix modelling, acoustic modelling, maximum mutual information

1. Introduction

Current language identification (LID) systems can be broadly divided into two major categories, namely the phonotactic and the Gaussian Mixture Model (GMM) and/or Support Vector Machine (SVM) based systems [1]. This paper focuses on the studies of the former approach. A typical phonotactic system consists of a phone recogniser (PR) as the front-end and a language model (LM) as the back-end. This type of system is commonly known as a PRLM (phone recogniser followed by language model) system. The front-end phone recogniser plays an important role of extracting the phonotactic information from the speech data. The performance of a PRLM system is greatly affected by the quality of the recogniser, as reported in [2].

Typically, the phone recogniser of a PRLM system is trained on speech data of a particular language using a phone set for that language. This may not be sufficient to cover the sound units that appear in the target languages of a language identification task. This problem may be circumvented by combining multiple PRLM systems trained on speech data from different languages with different phone sets to form a Parallel PRLM (PPRLM) system. PPRLM has been found to yield improvement over individual PRLM systems [1]. This is because different PRLM systems are capable of extracting complementary phonotactic information to provide greater diversification.

In this paper, an alternative methodology for building PPRLM systems is proposed where the diversification of the parallel systems is achieved by using different acoustic models trained on the same speech data with the same phone set. The purpose is to train complementary systems in terms of the errors made by the phone recognisers. Analogous to system combination for speech recognition, merging outputs from multiple systems with different error patterns will help improve the final performance. This paper examines the use of the structured precision matrix modelling techniques and the discriminative training paradigm to train different acoustic models for PPRLM.

Recently, structured precision (inverse covariance) matrix modelling techniques have been found to yield improved performance for speech recognition. In particular, the Semi-tied Covariance (STC) [3], and Subspace for Precision and Mean (SPAM) [4] models were successfully applied to the large vocabulary continuous speech recognition. Furthermore, discriminative training methods such as Maximum Mutual Information (MMI) [5] has also been shown to outperform the conventional Maximum Likelihood (ML) training approach. This paper will investigate the use of the above techniques to obtain different acoustic models for PPRLM language identification. The remaining of this paper is organised as follows: Sections 2 describes the phonotactic LID system and Section 3 introduces the proposed PPRLM system using alternative acoustic models. The formulation of the structured precision matrix models and MMI training are discussed in Sections 4 and 5 respectively. Experimental results are presented in Section 6.

2. Phonotactic Language Identification

The spoken language identification (LID) task is to recognise the language associated with a segment of speech data. One of the most commonly employed techniques for LID is the phonotactic approach. Phonotactic constraints are language specific in that the permissible combination of phonemes is unique to a language and it provides a good source of information for language recognition.

PRLM is a phonotactic system which consists of two parts: 1) the front-end of the system is a phone recogniser, responsible for extracting the phonotactic information (phone sequences) from speech data; 2) the back-end of the system is a language model for capturing the phonotactic constraints for each target language. As reported in [2], good quality phone recognisers are important to achieve a good PRLM system. It allows accurate and reliable extraction of phonotactic information.

Parallel PRLM (or PPRLM) is an extension to the PRLM system. A PPRLM system combines multiple PRLM systems together. This provides a better diversification to the system since different PRLM systems are capable of modelling different phonotactic constraints. The standard approach combines multiple PRLM systems whose front-end phone recognisers are trained on speech data from different languages with different phone sets. Since the individual systems are trained on different languages, they capture different acoustic characteristics

(and hence phonotactic attributes) from the speech data. Therefore, combining these systems together improves the overall language recognition performance. In general, the performance gain increases with a greater number of parallel systems.

3. Contrastive Acoustic Models for PPRLM

As previously mentioned, standard PPRLM systems rely on having multiple training corpora of different languages to provide a good diversification. This papers considers a different approach for building PPRLM systems without the need of having many different training corpora. Instead, different acoustic models are trained on the same speech corpora, but using different modelling techniques and training paradigms to form the contrastive parallel systems. This approach provides more parallel systems without requiring additional phonetically transcribed speech data.

Such an approach can also be viewed as an attempt to overcome the errors made by the phone recognisers. Phone recognisers used in the PRLM systems are not perfect. Errors introduced by the phone recognisers will inevitably degrade the performance of the back-end phonotactic language models. Different acoustic models may have different error patterns which are complementary across different systems. Hence, merging these individual systems may help recover some of the errors made by the individual systems. It is also believed that the language of the training data and the type of acoustic models provide different aspects of diversification. Thus, combining the two factors may further improve the PPRLM system performance. In this paper, different structured precision matrix modelling techniques and model training paradigms are used to train different acoustic models. These techniques will be discussed in the following two sections.

4. Structured Precision Matrix Modelling

A structured precision matrix modelling technique aims at achieving a compact model representation to model the correlations between feature elements, while maintaining efficient computation [6]. Examples of this technique, which have been successfully applied to speech recognition, include the Semi-tied Covariance (STC) [3] and Subspace for Precision and Mean (SPAM) [4] models. These methods will be briefly described next.

Semi-tied Covariance (STC) [3] models the precision matrix as follows:

$$\boldsymbol{P}^{(m)} = \sum_{i=1}^{d} \lambda_i^{(m)} \boldsymbol{S}_i = \sum_{i=1}^{d} \lambda_i^{(m)} \boldsymbol{a}_i \boldsymbol{a}_i' \tag{1}$$

where $\boldsymbol{S}_i = \boldsymbol{a}_i \boldsymbol{a}_i'$ are the rank one basis matrices and $\lambda_i^{(m)}$ are the corresponding basis coefficients for the mth Gaussian component in the system. The basis vector, $\boldsymbol{a}_i$, is a $d \times 1$ column vector, where d is the feature dimension. To ensure that $\boldsymbol{P}^{(m)}$ is a positive definite symmetric matrix, it is necessary that $\lambda_i^{(m)} > 0$ for all i. The update formulae for the basis coefficients and the basis vectors are given by

$$\lambda_i^{(m)} = \frac{1}{\boldsymbol{a}_i' \boldsymbol{W}^{(m)} \boldsymbol{a}_i} \quad \text{and} \quad \boldsymbol{a}_i' = \boldsymbol{c}_i' \boldsymbol{G}_i^{-1} \sqrt{\frac{\beta}{\boldsymbol{c}_i' \boldsymbol{G}_i^{-1} \boldsymbol{c}_i}} \tag{2}$$

where the sufficient statistics are given by

$$\boldsymbol{W}^{(m)} = \sum_{t=1}^{T} \gamma_t^{(m)} \left(\boldsymbol{o}_t - \boldsymbol{\mu}^{(m)} \right) \left(\boldsymbol{o}_t - \boldsymbol{\mu}^{(m)} \right)' \tag{3}$$

$$\boldsymbol{G}_i = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t^{(m)} \lambda_i^{(m)} \boldsymbol{W}^{(m)} \tag{4}$$

$$\beta = \sum_{m=1}^{M} \sum_{t=1}^{T} \gamma_t^{(m)} \tag{5}$$

$\gamma_t^{(m)}$ denotes the posterior probability of the mth Gaussian component given the observation vector $\boldsymbol{o}_t$ at time t. $\boldsymbol{c}_i$ is the column vector of cofactors of $\boldsymbol{A}$ corresponding to the ith row where $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_d \end{bmatrix}'$. The update of the basis coefficients and basis vectors are interleaved using formulae in equation (2).

A more general form of precision matrix model is the SPAM model [4]. This paper considers only modelling the subspace for the precision matrix alone [6]. The precision matrix is given by

$$\boldsymbol{P}^{(m)} = \lambda_0^{(m)} \boldsymbol{S}_0 + \sum_{i=1}^{n-1} \lambda_i^{(m)} \boldsymbol{S}_i \tag{6}$$

$\boldsymbol{S}_0$ is a special basis matrix, initialised as the average precision matrix of the system to yield a positive-definite symmetric matrix. Positive-definiteness of $\boldsymbol{P}^{(m)}$ can then be guaranteed by initialising $\lambda_0^{(m)} = 1$ and $\lambda_i^{(m)} = 0$ for $i \neq 0$.

The remaining basis matrices, $\boldsymbol{S}_i$ for $i = 1, 2, \ldots, n-1$, are initialised as the top $n-1$ symmetric matrices that span the space of $\{\boldsymbol{W}^{(m)}\}$ using Singular Value Decomposition (SVD). See [6] for more details. The basis matrices are initialised and kept constant for the remaining learning process. The basis coefficients are updated using a simple line search along the gradient of the objective function at the current parameter estimate. The detailed implementation is described in [6].

5. Maximum Mutual Information (MMI)

Recently, Maximum Mutual Information (MMI) [5] discriminative training paradigm has been found to outperform the conventional Maximum Likelihood (ML) training approach in speech recognition and other classification tasks. The ML objective function is given by

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{r=1}^{R} \log p(\mathcal{O}_r | \boldsymbol{\theta}) \tag{7}$$

where $\boldsymbol{\theta}$ is the model parameter set and $\mathcal{O}_r$ is the rth observation sequence. R denote the total number of training utterances. The ML estimation maximises the likelihood of each model generating the training data independently.

MMI training approach estimates the model parameters in a discriminative manner by maximising the following objective function

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{r=1}^{R} \log \left(\frac{p(\mathcal{O}_r, h_r^{(0)} | \boldsymbol{\theta})}{\sum_{s=1}^{S} p(\mathcal{O}_r, h_r^{(s)} | \boldsymbol{\theta})} \right) \tag{8}$$

where $h_r^{(0)}$ is the reference transcription of the rth utterance and $h_r^{(s)}$ denotes the competing hypotheses (confusions) generated

by the current model. The MMI parameter estimation formulae for the standard HMM systems and various precision matrix models can be derived using a weak-sense auxiliary function [6]. The resulting formulae are similar to the ML case, with the sufficient statistics replaced by the corresponding discriminative statistics [6].

6. Experimental Results

For preliminary experimentation, the German and Spanish data from the OGI Multilingual database [7] were chosen to demonstrate the usefulness of the contrastive acoustic model based PPRLM systems for LID. Each language has approximately 1.2 and 0.1 hours of training and test data respectively. A phone recongiser was trained for each language using a 3-state left-to-right HMM for each monophone. The state output distribution is modelled by a GMM with 8 components. 12 MFCC coefficients plus energy with the first and second derivatives form a 39-dimensional feature. Trigram language models were trained, one for each target language, for the back-end. The detection score for each trial is given by the posterior probability:

$$P(L|\mathcal{O}_r) = \frac{\mathcal{L}(\mathcal{O}_r|L)}{\sum_{\forall M} \mathcal{L}(\mathcal{O}_r|M)} \quad (9)$$

where $\mathcal{L}(\mathcal{O}_r|L)$ denotes the average likelihood of the rth test utterance, $\mathcal{O}_r$, given the language model, L. Fusion is performed by simply taking the sum of the scores of all the individual systems. The LID experiments are conducted as a language detection task and the performance is measured using the average Equal Error Rate (EER) of the target languages. Experimental results are reported on the 1996, 2003 and 2005 NIST Language Recognition Evaluation (LRE) sets with 30 seconds duration. There are 12 target languages for the 1996 and 2003 sets, and 8 for the 2005 set.[1].

6.1. Phone Accuracies

Table 1 shows the phone accuracies of the various German and Spanish phone recognisers on the 0.1 hours of test data. DIAGC refers to the standard diagonal covariance matrix structure. The baseline ML DIAGC system gave 26.79% and 37.77% phone accuracies for German and Spanish respectively. MMI training yielded approximately 10.0% absolute improvements for the DIAGC system for both languages. The STC system outperformed the DIAGC system by 5.0% and 1.0% for ML and MMI training respectively. The SPAM system gave the best performance with an absolute improvement of 7.0-8.0% over the DIAGC system for the ML case. Similar to the STC system, the gain for the MMI SPAM system is much smaller (0.8–1.1%).

Training Method	Precision Matrix	Phone Accuracies (%) German	Phone Accuracies (%) Spanish
ML	DIAGC	26.78	37.77
	STC	31.67	42.77
	SPAM	33.85	45.73
MMI	DIAGC	36.62	47.06
	STC	37.90	48.07
	SPAM	38.73	48.83

Table 1: Phone accuracies of the German and Spanish tokenisers using various acoustic modelling techniques

[1] Indian English was excluded from the 2005 test set due to lack of training data

6.2. Comparison of PRLM systems

The German and Spanish phone recognisers in Table 1 were used as the front-end tokeniser for the individual PRLM systems. The Equal Error Rate (EER) performance for these systems are tabulated in Table 2 for the 1996, 2003 and 2005 NIST Language Recognition Evaluation sets. Unlike the phone accuracy performance, there is no clear and consistent trend in the EER results across different acoustic modelling techniques. In most cases, the STC and SPAM models outperformed the DIAGC baseline by a small margin. However, no clear distinction can be drawn between the STC and SPAM models. Furthermore, the MMI method, which gave a much higher phone accuracy, turns out to be slightly inferior for PRLM language identification. This may be caused by the rather poor phone recognisers (less than 50.0% accuracy) such that the improvement from MMI training has negligible impact on the EER performance.

Tokeniser	Training Method	Precision Matrix	1996	2003	2005
German	ML	DIAGC	15.78	16.54	25.31
		STC	15.76	18.65	22.56
		SPAM	15.97	16.75	21.61
	MMI	DIAGC	16.33	18.44	21.16
		STC	16.87	18.83	22.07
		SPAM	17.26	18.14	22.65
Spanish	ML	DIAGC	15.48	17.45	22.75
		STC	15.48	16.78	22.98
		SPAM	14.44	16.28	22.29
	MMI	DIAGC	15.80	16.37	22.21
		STC	15.62	16.05	21.41
		SPAM	14.78	16.08	21.47

Table 2: Equal Error Rate (%) of various tokenisers on the 1996, 2003 and 2005 NIST Language Recognition Evaluation sets

6.3. Comparisons of Parallel PRLM (PPRLM) systems

Although the individual PRLM systems using different acoustic modelling techniques gave similar EER performance, the errors made by these systems may be different. Table 3 shows the EER performance comparison for three different 2-way PPRLM systems. `F1` is the standard PPRLM system combining two PRLM systems with different languages. `F2` and `F3` are PPRLM systems with different training paradigms and precision matrix modelling techniques respectively. All three PPRLM systems

Fusion System	Individual Systems	1996	2003	2005
F1	German+ML+DIAGC	15.78	16.54	25.31
	Spanish+ML+DIAGC	15.48	17.45	22.75
	Fusion	12.63	14.31	20.29
F2	German+ML+DIAGC	15.78	16.54	25.31
	German+MMI+DIAGC	16.33	18.44	21.16
	Fusion	13.56	15.34	19.93
F3	German+ML+DIAGC	15.78	16.54	25.31
	German+ML+SPAM	15.97	16.75	21.61
	Fusion	13.64	14.74	20.60

Table 3: Equal Error Rate (%) of various 2-way PPRLM systems on the 1996, 2003 and 2005 NIST LRE sets

gave consistent improvements over their respective individual systems. F1 gained 2.2–2.8% EER over the best individual systems. This is only slightly better compared to the gains from F2 and F3 (1.0–2.2%). Note that F2 and F3 were trained on only the German corpus and rely solely on alternative acoustic models to obtain EER reduction. In this case, the diversification given by alternative acoustic models is comparable to that provided by speech corpora of different languages.

Fusion System	Individual Systems	1996	2003	2005
F4	German+ML+ {DIAGC,STC,SPAM}	12.61	14.39	19.34
F5	German+MMI+ {DIAGC,STC,SPAM}	13.60	14.37	18.09
F6	F4 + F5	11.85	13.00	17.59
F7	Spanish+ML+ {DIAGC,STC,SPAM}	11.96	13.23	19.08
F8	Spanish+MMI+ {DIAGC,STC,SPAM}	12.53	13.28	19.08
F9	F7 + F8	10.91	12.22	18.02
F10	F6 + F9	9.64	10.90	15.67

Table 4: Equal Error Rate (%) of various PPRLM systems on the 1996, 2003 and 2005 NIST LRE sets

Table 4 shows the EER performance of more complex PPRLM systems. Systems F4, F5, F7 and F8 are 3-way PPRLM systems combining DIAGC, STC and SPAM models for different languages and training methods. F6 and F9 are 6-way PPRLM systems combining all the six acoustic models for German and Spanish respectively. It is evident from the results that increasing the number of parallel systems using different acoustic models consistently improve the LID performance, even if the contrastive systems were trained on the same speech corpus. Note that F4 to F9 are still single language PPRLM systems and the EER's for the 6-way PPRLM systems are about 0.8–2.7% lower than F1. Finally, combining all the individual systems together gives the 12-way F10 system with the lowest EER. The absolute gains were 3.0%, 3.4% and 4.6% on the 1996, 2003 and 2005 test sets respectively over the F1 system (without contrastive acoustic models).

Fusion System	Individual Systems	1996	2003	2005
F11	German+ML+DIAGC	15.78	16.54	25.31
	Spanish+ML+DIAGC	15.48	17.45	22.75
	Korean+ML+DIAGC†	18.59	19.31	21.41
	Japanese+ML+DIAGC†	31.13	32.22	35.28
	Fusion	12.91	13.76	19.81
F12	German+ML+DIAGC	15.78	16.54	25.31
	German+ML+SPAM	15.97	16.75	21.61
	Spanish+ML+DIAGC	15.48	17.45	22.75
	Spanish+ML+SPAM	14.44	16.28	22.29
	Fusion	10.67	12.81	17.26
F13	F11+F12	10.50	11.44	16.88

Table 5: Equal Error Rate (%) of various PPRLM systems on the 1996, 2003 and 2005 NIST LRE sets († denotes tokeniser trained on the IIR-LID data [8])

Lastly, two 4-way PRLM systems were compared in Table 5. F11 extends F1 by adding two systems of different languages (Japanese and Korean). These additional systems have EER's higher than those in F1 and only gave a small absolute gain of 0.5–0.6% on 2003 and 2005 test sets, compared with F1 (a 0.3% degradation was observed on the 1996 set). On the other hand, adding the German and Spanish ML SPAM systems to F1 reduced the EER by 1.5–3.0% absolute. The results show that it is preferable to include alternative acoustic models with similar EER performance than poorer PRLM systems of different languages. If all the six individual systems from F11 and F12 are combined to form F13, a further absolute EER reduction of 0.2–1.4% was obtained. This shows that the diversification due to different languages and acoustic models are complementary.

7. Conclusions

This paper has investigated the use of alternative acoustic modelling techniques for building a parallel PRLM (PPRLM) language identification system. Unlike the standard PPRLM systems where the individual systems are derived using phone recognisers of different languages and phone sets, the proposed method aims at providing more diverse systems that complement each other in terms of the errors made by the different phone recognisers of the same language. Preliminary experimental results show that different acoustic models is as important a factor as using training data from different languages in providing diversification to the PPRLM systems. Combining these two factors also showed promising improvements.

8. References

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[2] J. C. Pavel Matejka, Petr Schwarz and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proceedings Eurospeech*, September 2005.

[3] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[4] S. Axelrod, R. Gopinath, and P. Olsen, "Modeling with a subspace constraint on inverse covariance matrices," in *Proc. ICSLP*, 2002.

[5] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–48, Jan 2002.

[6] K. C. Sim and M. J. F. Gales, "Minimum phone error training of precision matrix models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 882–889, May 2006.

[7] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of the International Conference on Spoken Language Processing*, October 1992.

[8] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 2006.