

POSTERIOR FEATURES APPLIED TO SPEECH RECOGNITION TASKS WITH USER-DEFINED VOCABULARY

Guillermo Aradilla^{1,2}, Hervé Bourlard^{1,2}, Mathew Magimai-Doss¹

¹Idiap Research Institute, Martigny, Switzerland,

²École Polytechnique Fédérale de Lausanne (EPFL)

{aradilla,bourlard,mathew}@idiap.ch

ABSTRACT

This paper presents a novel approach for those applications where vocabulary is defined by a set of acoustic samples. In this approach, the acoustic samples are used as reference templates in a template matching framework. The features used to describe the reference templates and the test utterances are estimates of phoneme posterior probabilities. These posteriors are obtained from a MLP trained on an auxiliary database. Thus, the speech variability present in the features is reduced by applying the speech knowledge captured by the MLP on the auxiliary database. Moreover, information theoretic dissimilarity measures can be used as local distances between features.

When compared to state-of-the-art systems, this approach outperforms acoustic-based techniques and obtains comparable results to orthography-based methods. The proposed method can also be directly combined with other posterior-based HMM systems. This combination successfully exploits the complementarity between templates and parametric models.

Index Terms— Speech recognition, template matching, posterior features, Kullback-Leibler divergence

1. INTRODUCTION

Conventional automatic speech recognition (ASR) systems use the phonetic transcription to build the acoustic models of the system vocabulary. The phonetic transcription of each word is used to concatenate hidden Markov models (HMMs) representing phoneme-level units (typically context-dependent phonemes). In some applications the vocabulary is completely defined by the user and a proper phonetic transcription cannot be easily obtained. A typical example of this situation is a voiced-activated agenda consisting of user-defined names whose phonetic transcription may not appear in standard phonetic dictionaries. In this case, the user is then asked to provide a few acoustic samples or the orthography of each word.

Acoustic models in this type of applications can take different forms depending on the type of information (acoustic or orthographic) provided by the user to define the vocabulary. When words are characterized by a set of acoustic samples, template matching (TM) or HMMs can be applied. In the TM case, the acoustic samples are used as reference templates that are directly compared to the test utterances. The major advantage of this method is its simple implementation and its fast decoding time. However, its accuracy is strongly dependent on the pronunciation described by the templates and also, its performance can dramatically decrease when increasing the test vocabulary size. On the other hand, phoneme-level HMMs trained on a different (auxiliary) database can be used to form the

word models. The phonetic transcription required for this approach is then obtained by applying a phonetic HMM-based decoder to the acoustic sample provided by the user. A scheme of this method is shown in Figure 1(a). This latter approach outperforms the TM-based method because it applies the speech variability captured on the auxiliary database.

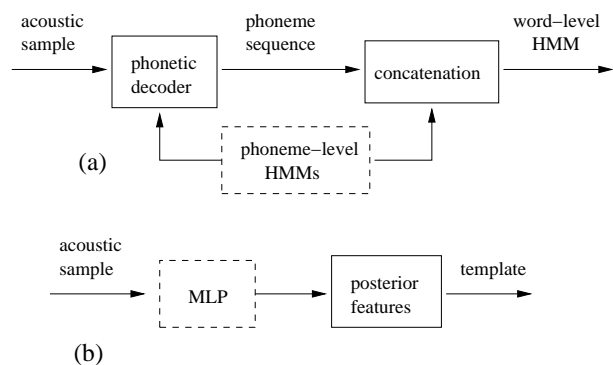


Fig. 1. Different paradigms when using an acoustic sample to define a word model. Figure (a) shows the implementation when using HMMs and Figure (b) shows the implementation proposed in this paper. Blocks in dashed lines are trained on an auxiliary database.

The orthography can also be used to infer the phonetic transcription in a HMM-based approach. In this case, the phoneme sequence appearing in Figure 1(a) is obtained from the orthography through a classification and regression tree (CART) [1]. This technique is widely applied in the speech synthesis field [2]. The major limitation of this approach appears on those words which do not follow the standard phonetic rules, like user-defined words.

In this paper, we present a novel approach when the vocabulary is defined by a set of acoustic samples. The presented approach is based on TM. The speech features forming the templates and the test utterances are estimates of phoneme posterior probabilities [3]. These posteriors are estimated by a multi-layer perceptron (MLP) which has been trained on an auxiliary database. This approach thus combines the advantages of both TM and HMM-based approaches because it benefits from a simple implementation and a fast decoding time and also, the speech variability from an auxiliary database can be incorporated through the posteriors at the feature level. Moreover, since phoneme posterior features can be seen as discrete probability distributions over the phoneme space, measures coming from the information theory field such as the entropy and the Kullback-Leibler (KL) divergence [4] can be successfully applied [5]. A scheme of

this approach is shown in Figure 1(b).

In addition, the presented method can be related to the KL-based acoustic model [6]. This model computes the KL divergence instead of the log-likelihood for estimating the state scores. Since both the HMM/KL model and the presented TM-based approach use the KL divergence as a local measure, they can be directly combined. In this work, we also show that the combination of these two methods can further improve the system accuracy.

This paper is structured as follows: Section 2 briefly describes posterior features. Section 3 summarizes the TM approach for ASR. Then, Section 4 presents the local distances that are used in this work. Section 5 describes the experiments and discusses the results and finally, Section 6 concludes this paper.

2. POSTERIOR FEATURES

Short-term spectral-based features, such as MFCC or PLP, are traditionally used in ASR. In this work, we use posterior-based features. Posterior probability of the phonemes given spectral features can be estimated by using a MLP [3]. This type of speech features have shown to be an efficient front-end for ASR because of their discriminative training and the ability of the MLP to model non-linear boundaries [7]. Moreover, the databases for training the MLP and for testing do not have to be the same so it is possible to train the MLP on a general-purpose database and use this posterior estimator to obtain features for more specific tasks as it has been shown in [8].

Given a sequence of spectral-based features $X = \{\mathbf{x}_1 \cdots \mathbf{x}_t \cdots \mathbf{x}_T\}$, Given two discrete probability distributions \mathbf{a} and \mathbf{b} of dimension a sequence of posterior features can be obtained $Z = \{\mathbf{z}_1 \cdots \mathbf{z}_t \cdots \mathbf{z}_T\}$. K . The KL divergence is defined as [4]: Each posterior feature \mathbf{z}_t is formed by concatenating the outputs of the MLP when using \mathbf{x}_t as input. Thus, $\mathbf{z}_t = [P(c_1|\mathbf{x}_t) \cdots P(c_k|\mathbf{x}_t) \cdots P(c_K|\mathbf{x}_t)]^\top$, where $\{c_k\}_{k=1}^K$ denotes the set of K phonemes¹. Since each posterior feature can be seen as a discrete probability distribution on the space of phonemes, dissimilarity measures over probabilities can be used to compare two frames. Section 4 describes the measures used in this work.

3. TEMPLATE MATCHING APPROACH

In TM for ASR, every word w in the lexicon \mathcal{W} is represented by a set of N_w samples $\mathcal{Y}(w) = \{Y_i(w)\}_{i=1}^{N_w}$ known as templates [9]. Each template $Y_i(w)$ is a sequence of speech features extracted from a particular pronunciation of w . When decoding a test sequence of features Z , a similarity measure $\varphi(Z, Y_i(w))$ is computed between the test sequence Z and each template $Y_i(w)$ of each word w of the lexicon \mathcal{W} . The test sequence Z is then decided to be the word \hat{w} associated to the template with the minimum distance.

$$\hat{w} = \arg \min_{w \in \mathcal{W}} \min_{Y \in \mathcal{Y}(w)} \varphi(Z, Y) \quad (1)$$

The choice of the similarity measure $\varphi(X, Y)$ is an important issue in this approach because it should take into account those properties from the templates that best describe the classes. In ASR, the most typical similarity measure is based on dynamic time warping (DTW) [9]. This measure handles the different speech rates that different pronunciations of the same word may have and it also uses a very similar decoding procedure than HMM.

When using DTW, a local distance between the speech vectors must be defined. Mahalanobis distance is typically used, although

¹In practice, a context of spectral features $\mathbf{x}_{t-4} \cdots \mathbf{x}_{t+4}$ is used as input for the MLP. Hence, each component of the posterior feature estimates $P(c_k|\mathbf{x}_{t-4} \cdots \mathbf{x}_{t+4})$.

previous experiments [5] have shown that the use of the KL divergence can yield better performance when using posterior features. In the next section, we describe the local distances used in this work.

4. LOCAL MEASURES

In this section, we describe the local measures for the DTW implementation used in this work. We use the Mahalanobis distance because it is the typical local distance used in TM and, since we use posterior features in this work, we also present several KL-based measures.

4.1. Mahalanobis Distance

Mahalanobis distance using diagonal covariance matrix is the most common similarity measure between features in a TM approach [9]. Given two feature frames \mathbf{a} and \mathbf{b} of dimension K , Mahalanobis distance is defined as

$$D_{Mahal}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^K w_i (a_i - b_i)^2 \quad (2)$$

where the weights w_i take into account the different variance associated to each component.

4.2. Kullback-Leibler Divergence

Mahalanobis distance using diagonal covariance matrix is the most common similarity measure between features in a TM approach [9]. Given two feature frames \mathbf{a} and \mathbf{b} of dimension K , Mahalanobis distance is defined as

$$KL(\mathbf{a}||\mathbf{b}) = \sum_{i=1}^K a_i \log \frac{a_i}{b_i} \quad (3)$$

This measure has its origin in the information theory field [4]. It defines the average number of extra bits that are used when coding an information source with distribution \mathbf{b} with a code that is optimal for a source with distribution \mathbf{a} . Given the asymmetric nature of the KL divergence, several formulations can be used as local similarity functions. Let us consider \mathbf{z} the frames corresponding to the test sequence and \mathbf{y} the frames from the templates.

- $D_{KL}(\mathbf{z}, \mathbf{y}) = KL(\mathbf{y}||\mathbf{z})$. In this case, frames from templates are considered as the reference distributions.
- $D_{RKL}(\mathbf{z}, \mathbf{y}) = KL(\mathbf{z}||\mathbf{y})$. Frames from the test sequence are considered as the reference distributions in this case.
- $D_{SKL}(\mathbf{z}, \mathbf{y}) = KL(\mathbf{y}||\mathbf{z}) + KL(\mathbf{z}||\mathbf{y})$. This is a symmetric version of the KL divergence. It has been successfully applied in other fields such as speech synthesis [10].
- Symmetric KL can be seen as a weighted sum between D_{KL} and D_{RKL} where weights are equal. In this paper, we also investigate the use of weights which are not uniform but dependent on the entropy of the distributions. This weighting strategy has been previously applied in the combination of posterior-based multi-stream ASR [11].

$$D_{weight}(\mathbf{z}, \mathbf{y}) = \frac{w_1}{w_1 + w_2} KL(\mathbf{y}||\mathbf{z}) + \frac{w_2}{w_1 + w_2} KL(\mathbf{z}||\mathbf{y}) \quad (4)$$

where the weights are inversely proportional to the entropy H of the distributions², i.e., $w_1 = \frac{1}{H(\mathbf{y})}$ and $w_2 = \frac{1}{H(\mathbf{z})}$. Since

²The singularity of this measure definition is avoided because posteriors estimated from a MLP are never zero.

the entropy is a measure of uncertainty, this measure weights each factor depending on the uncertainty of the reference distribution.

5. EXPERIMENTS AND RESULTS

5.1. Database Description

The database chosen for this work is Phonebook [12]. It is formed by utterances containing isolated words. The test part of this database consists of 8 subsets of 75 different words each. There are 12 realizations of each word. The first and the last utterances for each word are chosen as the first and the second acoustic sample. The result of each experiment is the average of the individual results obtained from each subset.

The additional database used to learn the speech variability is the Conversation Telephone Speech (CTS) database [13]. The utterances of this database consist of sentences pronounced by different speakers in telephone conversations. The MLP and HMMs representing context-dependent phonemes are trained on this database. When using HMM/GMM, 16 Gaussian distributions are used to describe each state emission likelihood.

5.2. Experimental Setup

The systems implemented in this work are mainly divided in three groups depending on if they use acoustic information, the orthography of the word or a combination of both.

5.2.1. Using the Acoustic Information

In this work, we carry out experiments using one or two acoustic samples. This represents that the user has pronounced each word of the test vocabulary once or twice.

System 1 : The word models are templates formed by PLP features [14]. This is the simplest system because no information from other databases is considered. Mahalanobis distance is used as local measure for the DTW implementation in this case.

System 2 : The word models are based on HMM/GMMs using PLP features. The phonetic transcription required to form the word models is obtained from a phonetic decoder applied to the acoustic sample. The phonetic decoder is also based on HMM/GMMs. One phonetic transcription is obtained from each acoustic sample. Hence, when two acoustic samples are used, each test word is described by two phonetic transcriptions. This method is described in Figure 1(a).

System 3 : This system implements the novel TM approach presented in this paper. As shown in Figure 1(b), posterior features obtained from a MLP are used to form the templates and the test utterances. In this case, Mahalanobis distance and all the KL-based measures described in Section 4 are used in the DTW implementation.

All the above systems use the acoustic information to build the word models. Systems 1 and 3 directly use the acoustic sample as a template and System 2 uses the acoustic sample to infer the phonetic transcription that will be used to form the HMM-based word model. Moreover, Systems 2 and 3 incorporate the information of the speech variability learned on the CTS database. While in System 2 this information is carried by the HMM/GMMs, in System 3 this information is applied by the MLP through the estimation of the posterior features.

5.2.2. Using the Orthography Information

The orthographic information used in this work is provided by the database. This information is used to infer the pronunciation transcription for each test word through a CART-based statistical model [1]. Then, word models based on phoneme-level HMMs can be built. Each word is only represented by one phonetic transcription because there is only one orthography for each word. Since the test words of the Phonebook database are common English words, we can expect accurate phonetic transcriptions.

System 4 : In this system, HMM/GMM is used to form the word models.

System 5 : Word models are formed by the HMM/KL acoustic model [6].

HMM/KL computes the KL divergence instead of the log-likelihood to estimate the local score. A complete description of HMM/KL can be found in [6]. The interest of using this type of model is that its score can be combined straightforward with other systems also using the KL divergence.

5.2.3. Using Orthographic and Acoustic Information

In this section, we describe a system that combines the information from both the orthography and the acoustic samples. Again, two results are provided for this system corresponding to the use of one or two acoustic samples.

System 6 : In this system, the scores given by the HMM/KL word model and the TM are combined. This is possible because the local measure used in both systems is the KL divergence. The combination strategy is the minimum score. Hence, the decoding criterion expressed in (1) is replaced by

$$\hat{w} = \arg \min_{w \in \mathcal{W}} \min \left\{ \left(\min_{Y \in \mathcal{Y}(w)} \varphi(X, Y) \right), S(w) \right\} \quad (5)$$

where $S(w)$ is the score given by the HMM/KL model for word w .

Other combination strategies such as the sum have also been experimented. However, the combination based on the minimum score has shown to be the best. This combination strategy has also been shown to be the best in other works [15].

System 6 benefits not only from the combination of two independent information sources (orthography and acoustic) but also from the complementarity of templates and a HMM-based parametric model.

5.3. Results

Table 1 shows the results obtained by the systems described above. The following conclusions can be drawn:

- As expected, System 1 yields the lowest performance because it does not incorporate information about the speech variability learned from an auxiliary database.
- Systems using the orthographic information generally yield a better accuracy than systems using the acoustic information. It must be noted that the phonetic transcription inferred from the orthography is particularly accurate because test words are common words. In applications with user-defined vocabulary, words would probably have a less accurate phonetic transcription and hence, it would yield a worse performance.

		Word Accuracy	
		1 sample	2 samples
System 1	D_{Mahal}	56.8	75.2
System 2		90.6	95.4
System 3	D_{Mahal}	81.5	88.4
	D_{KLL}	91.4	95.7
	D_{RKL}	90.1	94.0
	D_{SKL}	92.5	95.8
	D_{weight}	93.4	96.1
System 4		96.0	
System 5		94.7	
System 6	D_{weight}	96.4	97.2

Table 1. Word accuracy of the implemented systems. Systems using the acoustic information show two results corresponding to the use of one or two acoustic samples.

- The proposed method (System 3) outperforms the conventional TM approach (System 1). In addition, the use of KL-based local measures further increases the accuracy. In particular, D_{weight} yields improvement with respect to other measures because the contribution of D_{KLL} and D_{RKL} depends on the entropy of each distribution. Moreover, it can be observed that the accuracy of the proposed method when using D_{weight} is higher than state-of-the-art acoustic-based approach (System 2) and when using two templates, yields comparable results to state-of-the-art orthographic-based approach (System 4).
- The combination of the proposed method with HMM/KL further improves the accuracy of the system. This can be explained because both approaches are complementary in two ways. Firstly, word references are represented by two different type of models: templates and HMMs. Secondly, the information used to build these references is independent: templates are built from the acoustic information and HMM/KL is built from the orthographic information.

6. CONCLUSIONS

In this paper we have presented a novel approach for those applications where the vocabulary is defined by the user. This approach is based on TM where speech features are phoneme posterior estimates. In this paper, we confirm the suitability of applying KL divergence when using posterior features as observed in previous experiments [5]. We also show that a weighted combination of the KL divergence can further improve the accuracy.

In addition, this approach is related to posterior-based HMM systems [6]. Since both methods use KL divergence as local measure, they can be directly combined. We show that the combination further improves the accuracy of the system. This combination benefits from a double complementarity since (a) words are represented by templates, which can describe the dynamics of the trajectories generated by the speech features in fine detail, and HMMs, which have good generalization capabilities and (b) the information used to build the templates comes from the acoustic information whereas the HMMs are formed from the orthographic information.

7. ACKNOWLEDGMENTS

The authors would like to thank Joel Pinto and Dr. John Dines for their contribution to the MLP training process.

This work was supported by the EU 6th FWP IST integrated project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-033812). The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on “Interactive Multimodal Information Management (IM2)”.

8. REFERENCES

- [1] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series, 1984.
- [2] P. Taylor, A. W. Black, and R. Caley, “The Architecture of the Festival Speech Synthesis System,” *ESCA/OSCODA Workshop on Speech Synthesis*, pp. 147–152, 1998.
- [3] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, Boston, 1993, vol. 247.
- [4] T. M. Cover and J. A. Thomas, *Information Theory*. John Wiley, 1991.
- [5] G. Aradilla, J. Vepa, and H. Bourlard, “Using Posterior-Based Features in Template Matching for Speech Recognition,” *Proceedings of Interspeech*, 2006.
- [6] —, “An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features,” *Proceedings of ICASSP*, 2007.
- [7] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, “On Using MLP features in LVCSR,” *Proceedings of ICSLP*, 2004.
- [8] S. Sivasdas and H. Hermansky, “On the Use of Task Independent Training Data in Tandem Feature Extraction,” *Proceedings of ICASSP*, 2004.
- [9] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.
- [10] E. Klabbbers and R. Veldhuis, “Reducing Audible Spectral Discontinuities,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.
- [11] H. Misra, H. Bourlard, and V. Tyagi, “New Entropy Based Combination Rules in HMM/ANN Multi-stream ASR,” *Proceedings of ICASSP*, 2003.
- [12] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, “Phonebook: A Phonetically-rich Isolated-word Telephone-speech Database,” *Proceedings of ICASSP*, pp. 101–104, 1995.
- [13] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Ordelman, R. Ordelman, and S. Renals, “The Development of the AMI System for the Transcription of Speech in Meetings,” *Lecture Notes in Machine Learning for Multimodal Interaction*, pp. 344–356, 2005.
- [14] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech,” *The Journal of the Acoustic Society of America*, vol. 87, no. 4, 1990.
- [15] L. I. Kuncheva, “A Theoretical Study on Six Classifier Fusion Strategies,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.