POSTERIOR FEATURES FOR TEMPLATE-BASED ASR

Serena Soldo [†,‡], *Mathew Magimai.-Doss* [†], *Joel Pinto* [†], *Hervé Bourlard* [†,‡]

[†] Idiap Research Institute, Martigny, Switzerland
[‡]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{serena.soldo, mathew, joel.pinto, bourlard}@idiap.ch

ABSTRACT

This paper investigates the use of phoneme class conditional probabilities as features (posterior features) for template-based ASR. Using 75 words and 600 words task-independent and speaker-independent setup on Phonebook database, we investigate the use of different posterior distribution estimators, different distance measures that are better suited for posterior distributions, and different training data. The reported experiments clearly demonstrate that posterior features are always superior, and generalize better than other classical acoustic features (at the cost of training a posterior distribution estimator).

Index Terms— Speech recognition, template-based approach, posterior features.

1. INTRODUCTION

The two major paradigms to perform acoustic modelling in Automatic Speech Recognition (ASR) are statistical modelling (e.g. Hidden Markov Model, HMM) and template/instance based approach (e.g. by Dynamic Time Warping, DTW). Both these techniques aim to find the best match between acoustic input and a set of reference templates. HMMs model stochastic templates (sequence of states to model statistical distribution) thus providing high generalization properties, while DTW uses deterministic templates/instances (sequence of feature vectors) that take into account the details of the speech signal.

ASR systems typically use cepstral features. In such a case, the use of Euclidean distance for template-based approach can be interpreted as a special case of HMM-based approach, where each time frame in the reference template is a single state HMM with emission distribution modeled by a single Gaussian with a mean vector equal to the feature vector and unit covariance matrix. However, in practice, HMM-based approach uses mixture of Gaussians in order to provide a more flexible representation (to handle variability) and, thus, better performances compared to template-based approach (which may need large number of instances to handle variability).

Recently, the use of phoneme class conditional probabilities directly as speech features has been proposed for both HMM-based [1] and template-based ASR [2]. We refer to these features as *posterior features* (Section 2). In this case, the states of the HMM are modeled by a single multinomial distribution. The emission score is estimated by computing the Kullback-Leibler (KL) divergence between the state multinomial distribution and observation feature vector (i.e. posterior feature). The template-based approach also results in a comparison similar to HMM, where the posterior feature belonging to the reference template and the one belonging to the test template can be, again, compared through KL-divergence. From this, it can be observed that through the use of posterior features HMM-based approach and template-based approach can converge to a common framework. Such a common framework can be observed for spectral features in the case of learning vector quantization HMM [3].

This work focuses on template-based ASR using posterior features. In preliminary work, we used an MLP to estimate the posterior features. On small vocabulary task, it was shown that posterior features can yield significantly better performance than standard spectral-based features. However, there are alternate approaches to estimate posterior features. For instance, using Gaussians or GMMs (generative approach). As discussed later in the paper (Section 5), the choice of estimator can provide its own flexibility.

This paper builds on our previous work investigating the following three aspects. The first aspect investigates the use of GMM for estimation of posterior features and compares them to standard spectral feature. The second aspect is how does the posterior features (estimated by GMMs and MLPs) compare to standard spectral feature when the size of vocabulary is increased. Finally, the third aspect is the effect of the training data, such as the use of auxiliary/cross-domain data. This aspect is addressed by using *off-the-shelf* GMMs and MLPs, i.e. already-existing estimators trained on different data.

We investigate the above mentioned aspects on task-independent speaker-independent small vocabulary (75 words) and medium vocabulary (600 words) isolated word recognition tasks using Phonebook corpus as matched condition data, and with conversational telephone speech (CTS) corpus as auxiliary data. Our studies show that posterior features perform significantly better than the spectral based features on both small vocabulary task and medium vocabulary task irrespective of the type of posterior features estimator used (i.e. GMMs or MLP) and the type of data used (matched condition data or auxiliary data) to train the estimators.

Section 2 presents the different posterior feature estimators. In the preliminary works, we investigated the use of KL-divergence, Bhattacharya distance, and Euclidean distance as local distance measures for posterior features. In this work, we also investigate cosine distance and scalar product. We present the distance measures in Section 3. Section 4 describes the experimental setup and the results. Finally, Section 5 provides a discussion that puts this work into a broader context followed by conclusion.

2. POSTERIOR FEATURES

Robustness toward noise and speaker variability are two of the most challenging problems in ASR when standard spectral features are used. Traditional features, like MFCC or PLP, contain (desirable and undesirable) information that gives the feature space a high variability.

In template-based speech recognition, standard spectral features are usually transformed into a more stable representation [4]. In this work we use a method to transform standard spectral features into linguistically meaningful features, i.e. phoneme class conditional probabilities (or posterior features).

Formally, given a spectral-based feature vector at time t, x^t, and given a set of possible phoneme classes c_k with $k \in \{1, 2, ..., K\}$, the vector $\mathbf{p}^t$ of the posterior probabilities is given by $\mathbf{p}^t = [P(c_1|x^t), \ldots, P(c_K|x^t)] = [p_1 \cdots p_k \cdots p_K]$. As discrete distribution, the vector $\mathbf{p}^t$ has two properties: *a)* $\forall k \in \{1, 2, ..., K\}, \mathbf{p}_k^t \in [0, 1]$ and *b)* $\sum_{k=1}^{K} \mathbf{p}_k^t = 1$.

These probabilities can be directly estimated through the use of a well trained MLP [5]. Alternatively, posterior probabilities can be obtained through likelihood estimation (using Gaussian Mixture Model, GMM), according to Bayes' law:

$$P(c_k|x^t, \Theta) = \frac{p(x^t|c_k, \Theta) \cdot P(c_k|\Theta)}{p(x^t|\Theta)} \tag{1}$$

where $P(c_k)$ is the prior probability for the class c_k. In the reminder of the paper, we adopt a simpler representation by omitting the reference to the time, t, and to the parameter set, Θ.

2.1. GMM-based estimation

The likelihood of each phoneme class c_k can be modeled with a linear combination of Gaussian distributions (i.e. GMM):

$$p(x|c_k) = \sum_{n=0}^{N-1} \alpha_{n,k} \mathcal{N}(x|\mu_{n,k}, \Sigma_{n,k}) \tag{2}$$

where N is the number of Gaussian distributions, α_n are the mixing coefficients, $\mu_{n,k}$ and $\Sigma_{n,k}$ are the means and the variances of each Gaussian distribution for the class c_k, respectively.

The likelihood of the observation x given the class c_k, $p(x|c_k)$, computed with GMM, can be used to estimate posterior probabilities through Bayes' law:

$$P(c_k|x) = \frac{p(x|c_k) \cdot P(c_k)}{\sum_{c'_k} p(x|c'_k) P(c'_k)} \tag{3}$$

The parameters of GMM are trained maximizing the likelihood of the training data. Typically, this model provides adaptability and scalability but may also need large amount of training data to estimate the parameters robustly.

Recently, the posterior features estimated by GMM has been used for discriminative modeling in HMM-based ASR [6] as well as for keyword spotting [7] (in this case, using unsupervised training).

2.2. MLP-based estimation

Given an acoustic input vector x and a target phoneme class output c_k, a well trained MLP can directly estimate the class conditional probability $P(c_k|x)$ [5]. MLPs learn decision boundaries to discriminate optimally between models without making prior assumptions on the distribution of the data. The MLP training aims to minimize the classification error increasing the correct class probability while decreasing the wrong ones.

The discriminative training procedure of the MLP ensures that posterior features are more robust to noise than spectral-based features, while retaining the speech discriminatory information. Moreover, since little prior assumption is made on the distribution of the training data, several kinds of features can be used as input and different set of classes can be trained as output. Also, MLP training is scalable with more data. On the other hand, the convergence of the training algorithm of MLP is slower than GMM training.

3. DISTANCE MEASURES

Euclidean distance or Mahalanobis distance are commonly used to compare vectors of standard short-term spectral-based features. Let $\mathbf{x} = [x_1 \cdots x_m \cdots x_M]^T$ denote the spectral features vector belonging to the reference template and $\mathbf{y} = [y_1 \cdots y_m \cdots y_M]^T$ denote the spectral features vector belonging to the test template. The Euclidean distance between these two vectors is:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} (x_m - y_m)^2$$

When posterior features are used to represent the speech, Euclidean distance can be used simply replacing the spectral features vectors with the corresponding posterior features vector. However, by considering the probabilistic properties of posterior features, different distance measures can be used to compute local matching scores. Let $\mathbf{p} = [p_1 \cdots p_k \cdots p_K]^T$ denote the posterior feature vector that belongs to the reference template and $\mathbf{q} = [q_1 \cdots q_k \cdots q_K]^T$ denote the posterior feature vector that belongs to the test template. In this paper, we investigate the following distance measures:

1. *Weighted symmetric KL-divergence* ($wSKL$): KL-divergence is an asymmetric measure that computes the difference between two probability distributions. $wSKL$ is a symmetric version of KL-divergence which takes the uncertainty in the reference and test posterior features into account [2].

$$wSKL(\mathbf{p}, \mathbf{q}) = w_{\mathbf{p}} \cdot KL(\mathbf{p}, \mathbf{q}) + w_{\mathbf{q}} \cdot RKL(\mathbf{p}, \mathbf{q})$$

where

$$\begin{aligned} KL(\mathbf{p}, \mathbf{q}) &= \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}, \\ RKL(\mathbf{p}, \mathbf{q}) &= \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k}, \\ w_{\mathbf{p}} &= \frac{\frac{1}{H(\mathbf{p})}}{(\frac{1}{H(\mathbf{p})} + \frac{1}{H(\mathbf{q})})} \\ w_{\mathbf{q}} &= \frac{\frac{1}{H(\mathbf{q})}}{(\frac{1}{H(\mathbf{p})} + \frac{1}{H(\mathbf{q})})} \end{aligned}$$

$H(\mathbf{p})$ is the entropy of $\mathbf{p}$ and $H(\mathbf{q})$ is the entropy of $\mathbf{q}$.

2. *Bhattacharya distance* ($Bhatt$): Measures the similarity of two probability distributions.

$$Bhatt(\mathbf{p}, \mathbf{q}) = -\log(\sum_{k=1}^{K} \sqrt{p_k \cdot q_k})$$

3. *Cosine angle* ($cosine$): In a previous work [8], it was shown that MLP-based posterior features belonging to different classes tend to be orthogonal, thus they can be modeled using the cosine angle as distance measure.

$$cosine(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}^T \mathbf{q}}{|\mathbf{p}||\mathbf{q}|}$$

4. *Scalar product* (SP): As $cosine$, also SP is a measure of the orthogonality of posterior features. Moreover, it has been shown that scalar product is the "optimal" estimation of the

probability that a pair of posterior features vectors $(\mathbf{p}, \mathbf{q})$ belong to the same class when an MLP is used as estimator [8].

$$SP(\mathbf{p}, \mathbf{q}) \quad = \mathbf{p}^T \mathbf{q} \quad = \sum_{k=1}^{K} p_k \cdot q_k$$

More recently, the scalar product has also been used as the (theoretically optimal) measure of the probability that two posterior distributions results from the same underlying phonetic event [7, 9]. Given its light computational cost, scalar product can be suitable in case of large dimensional space representation, such as the case in [6].

4. EXPERIMENTAL SETUP

In this section, we define the different tasks studied and the estimators used for posterior features extraction.

4.1. Task

We use the Phonebook speech corpus [10] (PB) for speaker-independent task-independent, small vocabulary isolated word recognition. The test set consists of 8 lists of utterances, each containing 75 words uttered on average by 11 or 12 speakers once. None of the speakers present in the training data appear in cross validation and testing data. There are 42 context-independent phonemes including silence. We present recognition studies on the following two tasks:

1. *75 words task*: Recognition is performed on 8 lists separately, using the respective 75 words lexicon. The performance is measured as average word error rate.
2. *600 words task*: Recognition is performed using a single lexicon consisting of the words from all the 8 lists.

We adopt the same framework as in [2], where two random utterances of each word were used as reference templates. We use the state-of-the-art hybrid HMM/MLP system reported in [11] as reference system. On the 75 words task and 600 words task, the HMM/MLP system yields word error rates (WERs) of 1.2% and 4.0%, respectively.

4.2. Posterior features extraction

In this section, we describe the posterior features extraction using two different estimators, namely GMM and MLP, each of them trained on two different datasets: PB corpus (matched data) and Conversation Telephone Speech (CTS) corpus (auxiliary data).

The training set of PB corpus consists of 6.7 hours of speech data represented using 39-dimensional PLP cepstral coefficient extracted every 10ms. Using this data we trained two estimators:

- GMM: we trained monophone HMM models. Each model, corresponding to one of the 42 classes, has 3 states and each emission probability is modeled with a mixture of 32 Gaussians. To estimate the posterior probability, first state level probability was estimated (42x3 values) and then marginalized to 42 phoneme class probabilities.
- MLP: we trained an MLP taking as input a vector of PLP features along with a temporal context of 90ms. The MLP has 800 hidden units and 42 output units, each corresponding to a phonemic class.

CTS corpus consists of continuous speech utterances from different speakers over a telephone channel. This corpus contains 45 context-independent phonemes including silence. In our work, we used *off-the-shelf* systems that were already trained on CTS:

- GMM: trained using 277 hours of speech from CTS training set. In this case the number of Gaussians used to model the emission probability is 16. For further details about this system, the reader may refer to [12].
- MLP: trained using 232 hours of speech (a subset of the 277 hours used for GMM) from CTS training set. This system estimates 45 posterior probabilities. For further details about this system, the reader may refer to [11].

4.3. Results

Tables 1 and 2 present the recognition results measured in terms of word error rate (WER) across different distance measures for 75 words task and 600 words task, respectively.

It can be observed that, when Euclidean distance is used, simply transforming PLP features into posterior features (irrespective of the estimator) allows to achieve significantly better performance. In addition, the use of other distance measures, that take into account the probabilistic nature of posterior features, provides a further significant improvement.

It can also be observed that MLP-based posterior features perform better than GMM-based posterior features. This can be attributed to the discriminative training used for MLPs.

Concerning the use of existing estimators trained on CTS data, MLP and GMM perform differently. MLP-based feature yields similar or better results compared to the use of PB training data. GMM-based features, on the contrary, perform significantly worse. This suggests that MLP-based posterior features could be efficiently estimated using existing systems, thus eliminating the need for training a new estimator on task specific data.

In all the systems described, the best performance was obtained using wSKL as distance measure.

Distance measure	PLP	Posterior Features			
		PB		CTS	
		GMM	MLP	GMM	MLP
Euclidean	24.8	7.9	3.8	20.4	3.6
wSKL	-	2.0	1.1	6.3	0.9
bhatt	-	3.1	1.6	7.5	0.9
cosine	-	2.7	1.5	9.4	1.2
SP	-	2.6	1.4	8.2	1.4

Table 1. Word Error Rate (WER) on 75 words task. The hybrid HMM/MLP system on this task yields 1.2% WER.

5. DISCUSSION AND CONCLUSION

One of the main issues in template-based ASR is how to achieve good generalization with fewer number of templates. This work attempts to show that one way to achieve this is to transform the spectral feature into posterior feature. In doing so, we may not be retaining the finer spectral details. However, if such details are relevant for ASR and can be estimated robustly, then it may be possible to learn them through the estimators, such as MLP, or find alternate approaches to integrate them. In addition, the use of posterior features for template-based ASR allows the use of additional knowledge

Distance measure	PLP	Posterior Features			
		PB		CTS	
		GMM	MLP	GMM	MLP
Euclidean	43.4	20.9	10.2	43.3	9.9
wSKL	-	6.8	3.4	18.6	2.8
bhatt	-	8.0	3.3	21.5	2.9
cosine	-	9.5	4.1	26.1	4.1
SP	-	8.3	4.4	23.2	5.3

Table 2. Word Error Rate (WER) on 600 words task. The hybrid HMM/MLP system on this task yields 4.0% WER.

sources, such as lexical knowledge, existing ASR systems, cross-domain data. In other words, we could leverage from the large body of existing work on GMMs and MLPs in the context of HMMs. We discuss a few below.

In our study, we see that the MLP approach consistently performs better than the GMM approach. This can be attributed to the discriminative learning which makes the MLP posterior features less susceptible to variations, such as speaker or environment. However, GMMs could be efficiently adapted in unsupervised manner on small amount of data to handle such variations.

The use of MLP and GMM lets also to explore different posterior feature representations. For instance, one could use articulatory features estimated by MLP or, in case of GMM, use the clustered states of a context-dependent system to estimate posterior features. Also, the ability to achieve language independence could be inherited through the use of articulatory features, "universal" phonemes, or subspace Gaussians method [13]. The implication of this can be further appreciated if we consider languages which have no lexical resources (they are just spoken) and limited acoustic data. In such a case, template-based approach using posterior features could not only help in developing ASR systems but also in generating lexical resources.

As discussed earlier, the HMM-based approach and template-based approach can converge to a common framework when using posterior features. This could be further put to use to combine HMM-based and template-based ASR for languages, such as Arabic where defining a consistent subword unit based pronunciation is difficult or a challenging task.

One of the shortcomings with template-based approach is how to generalize to unseen words. This typically needs subword based approach. Finding reliable and fewer subword templates can be difficult with spectral-based features. However, in the case of posterior features there are different possibilities. For example, it may be possible to extract a fewer robust templates of subword units using KL-HMM (by alignment). Otherwise, exploiting the robustness provided by MLP-based posteriors, we could find and use exemplars/instances available through resources such as the web [1].

In conclusion, in the context of template-based ASR, this study through the use of (a) different posterior feature estimators trained on matched condition data and auxiliary data, (b) different vocabulary sizes, and (c) different distance measures clearly demonstrated that posterior features (at the cost of training an estimator) can yield significantly better performance than standard acoustic features. Thus, indicating that posterior features are promising alternative to spectral features in further pursuing template-based ASR research.

[1] http://www.forvo.com

6. ACKNOWLEDGEMENT

This work was partly supported by the European Union under the Marie-Curie Training project SCALE, Speech Communication with Adaptive LEarning and the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

7. REFERENCES

[1] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-based Acoustic Models in a Large Vocabulary Recognition Task," in *Proceedings of Interspeech, ISCA, Brisbane, Australia*, 2008.

[2] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Posterior Features Applied to Speech Recognition Tasks with User-Defined Vocabulary," in *Proceedings of ICASSP, Taipei, Taiwan*, 2009, pp. 3809–3812.

[3] H. Iwamida, S. Katagiri, and E. McDermott, "Speaker-independent Large Vocabulary Word Recognition Using an LVQ/HMM hybrid algorithm," in *Proceedings of ICASSP, Toronto, Ontario, Canada*, 1991, pp. 553–556.

[4] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-Based Continuous Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1377–1390, 2007.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[6] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proceedings of ICASSP, Philadelphia, PA, USA*, 2005.

[7] Y. Zhang and J. Glass, "Towards Multi-Speaker Unsupervised Speech Pattern Discovery," in *Proceedings of ICASSP, Dallas, Texas, USA*, 2010, pp. 4366–4369.

[8] A. Asaei, B. Picart, and H. Bourlard, "Analysis of Phone Posterior Feature Space Exploiting Class-Specific Sparsity and MLP-Based Similarity Measure," in *Proceedings of ICASSP, Dallas, Texas, USA*, 2010.

[9] T. Hazen, W. Shen, and C. White, "Query-by-example Spoken Term Detection using Phonetic Posteriorgram Templates," in *Proceedings of ASRU, Merano, Italy*, 2009, pp. 421–426.

[10] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: A Phonetically-rich Isolated-word Telephone-speech Database," *Proceedings of ICASSP, Detroit, Michigan, USA*, pp. 101–104, 1995.

[11] J. Pinto, M. Magimai.-Doss, and Hervé Bourlard, "MLP Based Hierarchical System for Task Adaptation in ASR," in *Proceedings of ASRU, Merano, Italy*, 2009.

[12] T. Hain, L. Burget, J. Dines, I. Mccowan, M. Lincoln, D. Moore, G. Garau, V. Wan, R. Ordelman, and S. Renals, "The Development of the AMI System for the Transcription of Speech in Meetings," in *in Proceedings of MLMI, Edinburgh, UK*, 2005.

[13] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, K. Nagendra Goel, M. Karafit, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," in *Proceedings of ICASSP, Dallas, Texas, USA*, 2010, pp. 4330–4333.