IDIAP RESEARCH REPORT

ON MLP-BASED POSTERIOR FEATURES FOR TEMPLATE-BASED ASR

Serena Soldo Mathew Magimai.-Doss
Joel Praveen Pinto Hervé Bourlard

Idiap-RR-37-2009

DECEMBER 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

ON MLP-BASED 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, jpinto, bourlard}@idiap.ch

ABSTRACT

We investigate the invariance of posterior features estimated using MLP trained on auxiliary corpus towards different data condition and different distance measures for matching posterior features in the context of template-based ASR. Through ASR studies on isolated word recognition task we show that posterior features estimated using MLP trained on auxiliary corpus with out any kind of adaptation can achieve comparable or better performance when compared to the case where the MLP is trained on the corpus same as that of the test set. We also show that local scores, weighted symmetric KL-divergence and Bhattacharya distance yield better systems compared to Hellinger distance, cosine angle, L1-norm, L2-norm, dot product, and cross entropy.

Index Terms— Posterior features, Automatic speech recognition, Templates, Local scores, Multilayer perceptron

1. INTRODUCTION

There is a renewed/growing interest in template-based automatic speech recognition (ASR) system for different reasons: (a) availability of large amount of data so as to better handle the undesirable variabilities (differences between speakers, accents, conditions etc.) in the templates in conjunction with the availability of large storage and computation resources; (b) perceptual studies suggests that human tends to store both verbal (spoken message related) and non-verbal (e.g. speaker, dialect) information as episodes/traces, and uses both these information during recognition [1, 2]. An episode can be likened to a template; (c) it has been found that combination of template-based ASR and hidden Markov model (HMM) based ASR can yield improved performance [3, 4].

In template-based ASR system [5], each speech unit (e.g., word) is represented by a set of reference templates. A template typically being a sequence of feature vectors for an utterance of the speech unit. The training phase consists of generation of the reference templates. The test phase consists of generation of test template, (optionally) reference templates selection, and search for the best matching word sequence (by matching against the reference templates). The choice of feature vector and distance measure for local score not only influences the performance of the system but also practical issues such as, storage space and computation time.

Earlier research on template-based ASR typically used standard short-term spectrum based features, such as cepstral coefficients as features and Euclidean/Mahalanobis distance as the local score. However, the spectral-based features can be susceptible to undesirable variabilities such as, speaker, environment etc. Thus, putting demand for larger number of reference templates to achieve better generalization, and for storage and computational resources. Recent works have focussed on transforming the standard spectral-based features to discriminative features that tend to carry more linguistic class related information [6, 7, 8, 9]. Particularly, in a more recent work the use of phoneme class conditional posterior probabilities estimated using multilayer perceptron (MLP) as features was proposed [8]. These features also referred to as posterior features benefit from the ability of a well trained MLP to achieve invariance towards speaker and environmental characteristics while transforming the input spectral-based feature vector to linguistically meaningful dimensions. Template-based ASR studies using the posterior features showed that they can yield better performance compared to standard spectral-based features using only a fewer number of templates. Furthermore, it was also found that local scores that take into account the probabilistic nature of the feature, such as Kullback-Leibler (KL) divergence, Bhattacharya distance yield better performance compared to geometric distance measures such as Euclidean distance (L2-norm) [10, 9].

This paper builds up on the previous work of using MLP-based posterior features for template-based ASR investigating the following two aspects:

1. Choice of MLP training data: One of the main requirements in using posterior features is the availability of a trained MLP. The MLP can be trained on the same corpus or an auxiliary corpus. In the previous studies [8, 9], the improvements over the spectral-based features have been observed for both the cases, i.e., using MLP trained on the same corpus as well as on an auxiliary corpus. However, there remains a question how invariant are posterior features estimated with MLP trained on an auxiliary corpus towards data condition. In other words, without any kind of adaptation is it possible to achieve same level of performance as the ideal case where the MLP is trained on the same corpus as the test set. This can be possibly answered by evaluating the posterior features estimated by MLP trained on auxiliary corpus in terms of amount of auxiliary data, local scores, and number of templates, and comparing against the ideal case where posterior features are estimated using MLP trained on the same corpus as the test set.
2. Choice of distance measure for local score estimation: Each dimension of posterior feature vector has physical significance, i.e., each dimension corresponds to a particular phoneme. Also, the posterior feature vector has properties such as each dimension can take only a value between 0.0 and 1.0 and, the sum over the dimensions is 1.0. So, it is possible to use local measures that explicitly takes into account the phoneme class information. Also, there are other geometric distance measures, such as, L1-norm and cosine angle, and probabilistic distance measures such as Hellinger distance that may also suit the posterior features.

We investigate these aspects on small vocabulary (75 words) and

medium vocabulary (600 words) isolated word recognition tasks using Phonebook corpus with conversation telephone speech corpus as the auxiliary corpus.

Section 2 presents the different local scores studied. Section 3 describes the experimental studies and results. Finally, in Section 4 we summarize the work.

2. LOCAL SCORE

Given a pair of K-dimensional posterior feature vectors $\mathbf{p} = [p_1 \cdots p_k \cdots p_K]^T$ corresponding to the reference template and $\mathbf{q} = [q_1 \cdots q_k \cdots q_K]^T$ corresponding to the test template different types of measures for local scores can be motivated:

- Geometric measure: In this case each posterior feature is treated like a higher dimensional vector and traditional distance metrics such as,
  1. Euclidean ($Eucl$)
  $$Eucl(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{K}(p_k - q_k)^2$$
  2. L1-norm ($L1\text{-}norm$)
  $$L1\text{-}norm(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{K}|p_k - q_k|$$
  3. Cosine angle ($cosine$)
  $$cosine(\mathbf{p},\mathbf{q}) = \frac{\mathbf{p}^T\mathbf{q}}{|\mathbf{p}||\mathbf{q}|}$$

  are estimated. It is interesting to note that the MLP parameters are typically learned by minimizing the cross entropy between one-hot-encoding target vector (i.e., 1.0 for true class and 0.0 for other classes) and the posterior probability vector at the output of the MLP. So in case of a well trained MLP, the output of the MLP can be expected to lie in the simplex of the posterior feature space. In case of $cosine$ distance an additional $\log$ operation can better take this aspect into account.
- Probabilistic measure: Here each of the posterior feature vector is treated as a discrete probability distribution and local score is estimated by using distance/divergence measures,
  1. Kullback-Leibler divergence
     (a) With $\mathbf{p}$ as reference distribution (KL)
     $$\begin{aligned} KL(\mathbf{p},\mathbf{q}) &= \sum_{k=1}^{K} p_k \log\frac{p_k}{q_k} \\ &= H(\mathbf{p},\mathbf{q}) - H(\mathbf{p}) \end{aligned}$$
     where, $H(\mathbf{p},\mathbf{q})$ is the cross entropy with $\mathbf{p}$ as the reference distribution and $H(\mathbf{p})$ is the entropy of distribution $\mathbf{p}$.
     (b) With $\mathbf{q}$ as reference distribution (RKL)
     $$\begin{aligned} RKL(\mathbf{p},\mathbf{q}) &= \sum_{k=1}^{K} q_k \log\frac{q_k}{p_k} \\ &= H(\mathbf{q},\mathbf{p}) - H(\mathbf{q}) \end{aligned}$$
     (c) Symmetric measure (SKL)
     $$SKL(\mathbf{p},\mathbf{q}) = KL(\mathbf{p},\mathbf{q}) + RKL(\mathbf{p},\mathbf{q})$$
     (d) Weighted symmetric measure ($wSKL$) [9]
     $$wSKL(\mathbf{p},\mathbf{q}) = w_{\mathbf{p}}\cdot KL(\mathbf{p},\mathbf{q}) + w_{\mathbf{q}}\cdot RKL(\mathbf{p},\mathbf{q})$$
     $$w_{\mathbf{p}} = \frac{\frac{1}{H(\mathbf{p})}}{(\frac{1}{H(\mathbf{p})}+\frac{1}{H(\mathbf{q})})}, \quad w_{\mathbf{q}} = \frac{\frac{1}{H(\mathbf{q})}}{(\frac{1}{H(\mathbf{p})}+\frac{1}{H(\mathbf{q})})}$$
  2. Bhattacharya distance ($Bhatt$)
  $$Bhatt(\mathbf{p},\mathbf{q}) = -\log(\sum_{k=1}^{K}\sqrt{p_k\cdot q_k})$$
  3. Hellinger distance ($Hellinger$)
  $$Hellinger(\mathbf{p},\mathbf{q}) = 1.0 - \sum_{k=1}^{K}\sqrt{p_k\cdot q_k}$$
- Linguistic measure: Here the measure is defined such that the local score takes into account the probability mass associated to each dimension (phoneme class) explicitly. In other words, measures how much a given pair of vectors belong to the "same phoneme class", such as,
  1. Dot product ($dotProd$)
  $$dotProd(\mathbf{p},\mathbf{q}) = \mathbf{p}^T\mathbf{q} = \sum_{k=1}^{K} p_k\cdot q_k$$
  The probability that pair of posterior feature vectors $(\mathbf{p},\mathbf{q})$ belong to the same class [11, 12].
  2. Cross entropy
     (a) $\mathbf{p}$ as reference distribution ($cross$)
     $$H(\mathbf{p},\mathbf{q}) = -\sum_{k=1}^{K} p_k\cdot\log(q_k)$$
     (b) $\mathbf{q}$ as reference distribution ($R\text{-}cross$)
     $$H(\mathbf{q},\mathbf{p}) = -\sum_{k=1}^{K} q_k\cdot\log(p_k)$$
     (c) Symmetric cross entropy ($S\text{-}cross$)
     $$S\text{-}cross(\mathbf{p},\mathbf{q}) = H(\mathbf{p},\mathbf{q}) + H(\mathbf{q},\mathbf{p})$$
     (d) Weighted symmetric cross entropy ($wS\text{-}cross$)
     $$wS\text{-}cross(\mathbf{p},\mathbf{q}) = w_{\mathbf{p}}\cdot H(\mathbf{p},\mathbf{q}) + w_{\mathbf{q}}\cdot H(\mathbf{q},\mathbf{p})$$

  These local scores have their counter parts in the local scores defined earlier in geometric sense and probabilistic sense. For $dotProd$, we use an additional $\log$ operation similar to $cosine$. We found it to be beneficial. It can be observed that RKL and $R\text{-}cross$ will yield exactly same results for isolated word recognition task as $H(\mathbf{q})$ is constant at each time frame across all the words.

In the previous work [8, 10, 9], local scores $Eucl$, KL, RKL, SKL, $wSKL$, and $Bhatt$ have been investigated but not all of them jointly on the same task. In this paper, we study the local scores described in this section together.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We perform isolated word recognition studies using Phonebook (PB) speech corpus [13]. The test set contains 8 different sub-sets of 75 different words spoken only once on an average by 11 different speakers. This setup was originally defined for speaker-independent task-independent HMM-based ASR [14] and, later adopted for template-based ASR with a two template and one template scenario in [9]. In this study, we use the same template-based ASR setup. Furthermore, in addition to studies on the small vocabulary 75 words task we also perform studies on 600 words task (created by merging the 8 sets) as done in [14, 15].

For estimating posterior features,

- MLP trained on the same corpus: We use off-the-shelf MLP trained to classify context-independent phonemes on Phonebook corpus with 6.7 hours of speech data for speaker-independent task-independent HMM-based ASR system. The hybrid HMM/MLP system on the test set described above yield word error rates (WERs) of 1.2% and 4.0% for 75 words task and 600 words task, respectively [15].

- MLP trained on auxiliary corpus: We use off-the-shelf MLPs trained with varying amount of speech data on conversation telephone speech (CTS) corpus to classify context-independent phonemes.

For further details about the MLPs, the reader may refer to [16, 15]. One of the main strengths of template-based ASR is that the templates can be obtained from entirely different data set/condition than the one used to test. In this sense, generation of reference templates with posterior features estimated using MLP trained on Phonebook serves as an idealistic scenario, where except for different speakers and words present in the training and test data, a good match between training and test data conditions can be expected.

3.2. Results and Discussion

Tables 1 and 2 present the results measured in terms of WER across different local scores described earlier in Section 2 for 75 words task and 600 words task, respectively. The major observations are summarized as follows:

1. Posterior features estimated from MLP trained with "sufficient amount of data" on auxiliary corpus can yield performance comparable or better than the ideal well matched scenario i.e., posterior features estimated from MLP trained on the same corpus. This is observed for both two templates case and one template case.
2. For all local scores, the performance generally improves with the increase in the CTS MLP training data for both two templates case and one template case. The probabilistic local scores tend to achieve performance closer to the matched scenario with fairly low amount of MLP training data (compared to other local scores), especially on 600 words task. Overall, local score $wSKL$ consistently yields the best system for both 75 words and 600 words tasks with local score $Bhatt$ being close next best.
3. Among the geometrically motivated local scores, $cosine$ yields the best system for both the tasks. Interestingly, $L1\text{-}norm$ yields a system that performs better than the system using $Eucl$ as local score.
4. In the case of the local scores that take into account probabilistic nature of the feature, $wSKL$, $Bhatt$, and SKL perform better and yield competing systems. As the amount of CTS MLP training data increases KL and $Hellinger$ yield competing systems. Local score RKL yields the lowest performance.
5. $S\text{-}cross$ yields the best system among the linguistically motivated local scores. Furthermore, all the local scores i.e., $dotProd$, $cross$, $S\text{-}cross$, and $wS\text{-}cross$ yield performance lower than their counter parts, $cosine$, KL, SKL, and $wSKL$, respectively. An exception being $S\text{-}cross$ yielding a competitive system compared to SKL on 75 words task. The lower performance can be due to the following reason. Given two pairs of posterior vectors $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{x}, \mathbf{y})$, where, $\mathbf{p} = \mathbf{q}$, $\mathbf{x} = \mathbf{y}$, and $\mathbf{p} \neq \mathbf{x}$ local scores such as $cosine$ or KL will yield exactly same score for both the pair of vectors but $dotProd$ or $cross$ will yield entirely different scores. It can be also observed that the gap in the performance typically reduces as the CTS MLP training data increases.

It can be observed that there is a wider performance gap between the systems using local scores $cross$ and KL, especially for one template case of 600 words task. The essential difference between the two local scores is the use of extra information $H(\mathbf{p})$ by KL. If the reference template feature vector $\mathbf{p}$ were to be a delta (δ) distribution (i.e., all probability mass assigned to one dimension), the local scores $cross$ and KL are same. Since we use the estimate of phoneme posterior probabilities as posterior feature the difference in the performances can possibly be explained in terms of mismatch between pronunciation model and observation where, the reference template serves as the pronunciation model. For instance, the sequence of phoneme posterior probability vectors in the reference template with $\arg\max$ operation over the dimensions of each vector can be likened to pronunciation models in standard HMM-based ASR system. We know each test template posterior feature vector contains the probability for each phoneme. The local score $cross$ can be now seen as similar to local score estimation in hybrid HMM/MLP system (assuming equal priors for all phoneme classes). However, the reference template is susceptible to errors made by MLP and thus can serve as an erroneous pronunciation model. This can lead to mismatch and yield lower performance. In case of KL, the problem of matching with erroneous pronunciation is handled by taking uncertainty in the model (which is different for different words at each time instant) into account via entropy $H(\mathbf{p})$. This is also indicated by almost 4% drop in the performance gap for 600 words task when the CTS MLP training data increases (which in turn can lead to better estimation of posterior features). In addition, it can be said that when using $cross/KL$ as local score we are introducing pronunciation model like constraints similar to standard HMM-based system.

Along the similar lines parallel between the system using $R\text{-}cross/RKL$ as local score and knowledge-based ASR approach [17] can be drawn. When estimating $R\text{-}cross/RKL$ the test template feature vector $\mathbf{q}$ is the reference distribution which is same at each time frame across all the words. So, each test template feature vector serves like a phoneme classification output or can be seen as segmentation of the test template into phonemes (each segment being of length 1 frame). As described earlier, the reference template can be seen like a pronunciation model. Given this, the estimation of $R\text{-}cross$ and the decoding process is equivalent to lexical matching or computation of a quantity like weighted Levenshtein distance. Alternatively, these interpretations about $cross/KL$ and $R\text{-}cross/RKL$ can also be simply visualized by converting the respective reference distribution to δ distribution by assigning all probability mass to the best phoneme class dimension (found by $\arg\max$ operation). Systems using local scores SKL, $wSKL$, $S\text{-}cross$, and $wS\text{-}cross$ can be seen as benefiting from the mix of two different methods i.e., statistical HMM-based ASR like pronunciation model constraint introduced by the use of local score $KL/cross$ and knowledge-based ASR like lexical matching introduced by the use of local score $RKL/R\text{-}cross$.

4. SUMMARY

In the context of template-based ASR, we investigated the invariance of posterior features towards data condition by using MLPs trained with varying amount of auxiliary data and comparing it against MLP trained on the same corpus as the test set. In conjunction with it we also studied different local measures. Our studies showed that posterior features estimated using MLP trained with sufficient amount of auxiliary data can achieve comparable or better performance than the MLP trained on the same corpus. In addition the studies showed that local scores, weighted symmetric Kullback Leibler divergence and Bhattacharya distance yield better systems. Furthermore, we also elucidated the use of local scores based on Kullback Leibler divergence and cross entropy in terms of standard HMM-based ASR like pronunciation model constraint and knowledge-based ASR like lexical matching. This aspect is open to future research.

	Corpus	CTS							PB	CTS							PB
	Hours	232	116	69	46	23	10	6	6.7	232	116	69	46	23	10	6	6.7
	#Template	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
Geometric	Cosine	1.2	1.1	1.4	1.3	1.3	1.9	1.8	1.4	1.8	2.0	2.3	2.1	2.4	3.0	3.6	2.4
	$L1\text{-}norm$	1.9	2.1	2.2	2.0	2.7	3.4	3.2	2.2	4.1	4.5	4.8	4.7	5.4	6.6	7.2	4.9
	$Eucl$	3.4	3.2	4.0	3.7	4.2	5.5	5.9	3.1	6.0	6.1	7.0	7.0	7.8	9.8	10.7	6.4
Probabilistic	$wSKL$	**0.9**	**0.9**	**0.9**	**0.9**	1.0	**1.1**	**1.2**	**1.1**	**1.4**	**1.4**	**1.6**	**1.6**	**1.7**	**2.2**	**2.6**	**2.0**
	SKL	**0.9**	1.0	1.0	1.1	1.0	1.3	1.5	1.3	1.6	1.7	1.9	1.8	2.0	2.5	3.0	2.3
	KL	**0.9**	1.1	1.0	1.2	1.0	1.5	1.4	1.3	1.9	2.1	2.1	2.4	2.3	3.3	3.6	3.0
	RKL	1.8	1.7	2.0	1.9	2.2	2.5	2.8	1.7	2.8	2.9	3.1	3.1	3.7	4.4	5.0	3.6
	$Bhatt$	**0.9**	**0.9**	1.0	1.0	1.1	1.3	1.6	**1.1**	1.5	1.7	1.8	1.7	2.0	2.5	3.0	2.1
	$Hellinger$	1.1	1.1	1.1	1.1	1.3	1.6	1.9	1.3	2.0	2.0	2.2	2.3	2.5	3.1	3.7	2.7
Linguistic	$dotProd$	1.4	1.6	1.5	1.3	1.8	2.1	2.4	1.4	2.2	2.4	2.5	2.4	2.8	3.6	4.1	2.3
	$cross$	2.5	2.8	2.7	2.9	3.1	4.3	4.8	2.1	5.2	5.6	6.3	6.2	6.7	8.9	9.5	4.8
	$R\text{-}cross$	1.8	1.7	2.0	1.9	2.2	2.5	2.8	1.7	2.8	2.9	3.1	3.1	3.7	4.4	5.0	3.6
	$S\text{-}cross$	**0.9**	1.1	1.0	1.1	**0.9**	1.3	1.5	**1.1**	1.5	1.7	1.8	1.8	1.9	2.7	3.1	2.3
	$wS\text{-}cross$	1.6	1.6	1.7	1.8	2.0	2.4	2.4	1.2	2.6	2.6	2.8	2.9	2.9	3.9	4.2	2.3

Table 1. WER averaged over 8 sub-sets on 75 words task. Boldface represents the best performing system across different local measures.

	Corpus	CTS							PB	CTS							PB
	Hours	232	116	69	46	23	10	6	6.7	232	116	69	46	23	10	6	6.7
	#Template	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
Geometric	$cosine$	4.1	4.1	4.6	4.6	5.3	6.9	8.0	4.1	7.1	7.2	7.8	8.1	9.3	11.3	13.0	8.2
	$L1\text{-}norm$	6.7	7.1	7.5	8.0	9.1	11.2	11.9	7.4	13.0	13.5	14.5	15.0	15.9	18.5	20.0	15.2
	$Eucl$	9.9	9.8	10.3	11.1	12.8	16.3	16.9	9.4	16.4	16.4	17.3	18.2	20.0	24.1	26.1	17.0
Probabilistic	$wSKL$	**2.8**	**2.9**	**3.1**	**3.5**	**3.8**	**4.9**	**5.8**	3.4	**5.7**	**6.0**	**6.1**	**6.7**	**7.5**	**9.4**	**10.7**	7.7
	SKL	3.0	3.3	3.5	3.6	4.0	5.2	6.2	3.7	6.3	6.7	6.9	7.4	8.2	10.2	11.5	8.4
	KL	3.6	3.7	4.2	4.1	5.0	6.2	6.6	4.2	7.7	7.9	8.9	8.9	9.8	11.5	12.8	9.8
	RKL	5.2	5.5	5.8	6.2	7.4	8.5	9.9	6.6	9.6	10.3	10.6	11.1	12.4	15.0	16.7	12.1
	$Bhatt$	2.9	3.2	3.3	3.7	4.1	5.2	6.4	**3.3**	6.0	6.4	6.5	7.0	8.1	9.6	11.3	**7.6**
	$Hellinger$	3.5	3.8	4.0	4.5	4.9	6.2	7.3	4.1	7.5	7.8	8.0	8.7	9.5	11.6	13.0	9.3
Linguistic	$dotProd$	5.3	5.7	5.9	6.3	6.9	8.5	9.6	4.4	8.5	8.9	9.5	10.0	10.8	13.1	14.8	8.5
	$cross$	9.8	10.6	10.8	11.8	12.4	15.1	17.4	8.2	17.7	18.2	19.6	20.5	21.2	24.6	26.7	16.4
	$R\text{-}cross$	5.2	5.5	5.8	6.2	7.4	8.5	9.9	6.6	9.6	10.3	10.6	11.1	12.4	15.0	16.7	12.1
	$S\text{-}cross$	3.4	3.8	4.0	4.2	4.6	5.6	6.7	3.9	6.9	7.3	7.6	8.2	8.7	10.6	12.3	8.9
	$wS\text{-}cross$	5.8	6.0	6.8	7.1	7.3	9.3	10.1	4.7	9.5	9.8	10.6	11.5	11.8	14.1	15.5	9.2

Table 2. WER for 600 words task. Boldface represents the best performing system across different local measures.

5. ACKNOWLEDGMENT

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

6. REFERENCES

[1] S. D. Goldinger, "Echoes of Echoes? An Episodic Theory of Lexical Access," *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998.

[2] H. Strik, "Speech is like a box of chocolates...," *Proceedings of 15th ICPhS*, pp. 227–230, 2003.

[3] S. Axelrod and B. Maison, "Combination of Hidden Markov Models with Dynamic Time Warping for Speech Recognition," *Proceedings of ICASSP*, 2004.

[4] G. Aradilla, J. Vepa, and H. Bourlard, "Improving Speech Recognition Using a Data-Driven Approach," *Proceedings of Interspeech*, pp. 3333–3336, 2005.

[5] L. R. Rabiner and H. W. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs New Jersey, 1993.

[6] K. Demuynck, J. Duchateau, and D. Van Compernolle, "Optimal Feature Sub-Space Selection Based on Discriminant Analysis," *Proceedings of Eurospeech*, pp. 1311–1314, 1999.

[7] 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.

[8] G. Aradilla, J. Vepa, and H. Bourlard, "Using Posterior-Based Features in Template Matching for Speech Recognition," *Proceedings of Interspeech*, 2006.

[9] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Posterior Features Applied to Speech Recognition Tasks with User-Defined Vocabulary," *Proceedings of ICASSP*, 2009.

[10] G. Aradilla and H. Bourlard, "Posterior-Based Features and Distances in Template Matching for Speech Recognition," *Proc. of MLMI*, 2007.

[11] B. Picart, "Improved Phone Posterior Estimation Through k-NN and MLP-Based Similarity," Tech. Rep. Idiap-RR-18-2009, 2009.

[12] A. Asaei, B. Picart, and H. Bourlard, "Analysis of Phone Posterior Feature Space Exploiting Class-Specific Sparsity and MLP-Based Similarity Measure," *Submitted to ICASSP 2010*.

[13] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: A Phonetically-rich Isolated-word Telephone-speech Database," *Proc. of ICASSP*, 1995.

[14] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN systems for training independent tasks: Experiments on 'PhoneBook' and related improvements," in *Proc. of ICASSP*, 1997.

[15] J. Pinto, M. Magimai.-Doss, and Hervé Bourlard, "MLP Based Hierarchical System for Task Adaptation in ASR," in *Proc. of ASRU (to appear)*, 2009 [Online]. Available http://www.idiap.ch/~jpinto/pubs/adaptation.pdf.

[16] J. Pinto, G.S.V.S Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP-based Hierarchical Posterior Estimator," *Accepted for IEEE Transactions on Audio, Speech and Language Processing*, 2009 [Online]. Available http://www.idiap.ch/~jpinto/pubs/hierarchy.pdf.

[17] D. H. Klatt, "Review of the ARPA speech understanding project," *J. Acoust. Soc. Amer.*, vol. 62, no. 6, pp. 1345–1366, 1977.