Interpolation of Acoustic Models for Speech Recognition

Thiago Fraga-Silva[1,2], *Jean-Luc Gauvain*[1], *Lori Lamel*[1]

[1] LIMSI - CNRS, B.P. 133, 91403 Orsay, France
[2] Université Paris-Sud, 91403 Orsay, France
{thfraga, gauvain, lamel} @limsi.fr

Abstract

Acoustic models for speech recognition are often trained on data coming from a variety of sources. The usual approach is to pool together all of the available training data, considering them all to be part of a unique training set. In this work, assuming that each source may have a different degree of relevance for a given target task, two techniques are proposed to weigh subsets of the training data. The first one is based on the interpolation of the model probability densities, while the other on data weighting. An method to automatically select the mixture coefficients is also proposed. The best technique presented here outperformed unsupervised MAP adaptation and led to improvements in word accuracy (up to 6% relative) over the pooled model.

Index Terms: Acoustic modeling, model interpolation, adaptation

1. Introduction

In Large Vocabulary Continuous Speech Recognition (LVCSR) systems, the acoustic phenomena is usually modeled by hidden Markov models (HMM). The parameters of such models are often estimated by maximizing the likelihood function on the training data [1]. The accuracy of the models estimated via the *maximum likelihood* (ML) criterion are strongly dependent on at least two factors. First, the ML estimation assumes that the training and the test sets are drawn from the same distribution function. In other words, the acoustic model will be applied to recognize some data that are similar to the training set. Second, the method relies on the use of a large amount of data to generate robust estimates.

The most common approach used for acoustic modeling is to pool together all of the available training data, considering them all to be part of a unique training set. This method leads to good accuracy if the two aforementioned assumptions hold. However, gathering a fair amount of acoustic training data that matches well the target is a hard task for many different domains, such as conversational telephone speech recognition [2] or non-native speech recognition [3, 4]. This problem is often treated using an adaptation method [4, 5, 6], such as *Maximum Likelihood Linear Regression* [7] or *Maximum a Posteriori* (MAP) [8, 9] adaptation. In such cases, an initial (and more general) well-trained model is adapted to some specific domain using a training subset that matches better the target task.

For some tasks, such as broadcast recognition, a large amount of acoustic data can be easily collected from a variety of sources. In such a case, pooling the training data usually leads to overall good performance levels on test sets that also come from a wide variety of sources [10]. However, it is reasonable to consider that each of the training sources can match differently a given test set or another. In general, some improvement of performance can be obtained by adapting the pooled model to some data coming from the same source of the target set [11]. However, the adaptation techniques allow similarity levels to be adjusted to only one of the sources. This work hypothesizes that better parameter estimates can be obtained by considering different degrees of relevance for each of the training sets.

In this direction, this paper proposes a method to measure these degrees of relevance, assigning a mixture coefficient to each training source. These coefficients are estimated using an Expectation-Maximization (EM) algorithm and a held-out data set with its associated manual or automatic transcriptions. In the same manner, two methods to take into account these coefficients are presented. They are both derived from the linear interpolation of the component models associated to each of the training sources. In the first method, the combination of the component models is performed only at the decoding phase. In the second method, the combination of the models is performed at the training phase and requires the re-estimation of the model parameters. It follows that the latter method is approximately equivalent to perform a data weighting during the model parameter estimation.

The principle applied here is strongly inspired by a similar approach widely used on language model training. Such models are commonly obtained by the interpolation of component language models, each one estimated from a different text source. The mixture coefficients are usually obtained by maximizing the likelihood function (or equivalently, minimizing the perplexity) on some held-out data.

Interpolation of acoustic models has been used in other tasks related to speech processing. For instance, for speech synthesis, it has been used to combine models of different speaking styles [12] or models trained for different speakers [13]. For speech recognition, model interpolation has been used to combine native and non-native acoustic models for a non-native speech recognition task as an alternative to acoustic model adaptation [4, 14]. However, in the previously report work, the mixture coefficients were manually selected and the model interpolation technique was assessed only at the decoding phase. This work shows that better performance levels can be achieved by considering the interpolation coefficients during model parameter estimation. Furthermore, it is shown that optimal coefficients (in the likelihood sense) can be automatically estimated.

This paper is organized as follows. Section 2 briefly describes acoustic model training and its use in the recognition system. In Section 3, we propose two methods to interpolate component acoustic models. Section 4 describes the method used to estimate the mixture coefficients. Section 5 describes the recognition system. In the Section 6, the experiments and results are discussed. A brief conclusion is given in Section 7.

This work has been partially supported by OSEO, the French State agency for innovation, under the Quaero program.

2. Acoustic models

In automatic speech recognition, the speech units are usually modeled by continuous density HMMs with Gaussian mixture state observation densities. Training these models requires an alignment between the audio stream and the associated phonemic representation. When manual reference transcriptions are available, the audio data is segmented in phones by a forced alignment procedure, using a pronunciation dictionary and an initial acoustic model. If no transcription is available, unsupervised training [15] can be applied. In this case, an initial system is used to decode the untranscribed data. The acoustic model parameter estimation is guided by one or many segmentation hypotheses given by the decoder. In this work, the acoustic models were trained using the unsupervised method with the best hypothesis given by the decoder as ground truth.

The HMM parameters and the state Gaussian mixture model (GMM) parameters can be obtained by a ML estimation procedure, usually defined as:

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}} f(X|\lambda) \tag{1}$$

where λ represents the model parameters, X the observed feature vectors and $f(.)$ the likelihood function.

Equation 1 is an incomplete data problem [16] for both, the HMM and GMM parameters. This problem can be solved iteratively using an EM algorithm. The focus of this paper is on the estimation of the GMM parameters, but can be extended to the HMM parameter estimation.

3. Acoustic model interpolation

The acoustic model estimation is commonly performed considering the training data as a single homogeneous data set. The ML estimation is performed over all the available data. At the decoding phase, the likelihood function $f(x|\lambda)$ of the estimated model λ is used together with a language and a pronunciation models in order to find the best possible sequence of words given an observed feature vector x.

The usual approach do not take into account the different degrees of relevance of each of the training sources. The following sections propose two methods to consider them. In this work, it was assumed that all the acoustic models have the same structure. Hence, the methods presented were used to interpolate the GMMs of each of the HMM states. It was also assumed that the training data can be split into K independent subsets, such as $X = \{X^1 \cdots X^K\}$. Each subset X^k contains a training vector with dimension T_k , represented by $X^k = \{x_1^k \cdots x_{T_k}^k\}$.

3.1. Gaussian mixture model interpolation

The first interpolation method proposed is straightforward. First, a model λ_k is estimated on each of the X^k subsets. The state GMMs are linearly interpolated at the decoding phase, assuming that each model contributes to the acoustic likelihood of an observed vector x_t with a coefficient α_k as follows:

$$f_{s_t}(x_t|\lambda) = \sum_{k=1}^{K} \alpha_k \cdot f_{s_t}(x_t|\lambda_k) \tag{2}$$

where $f_{s_t}(.)$ is the GMM density function of a state s_t at time t and $\sum \alpha_k = 1$.

This interpolation approach can be used in two different ways. First, one can perform the interpolation of the component models λ_k at runtime. It is also possible to build a model beforehand, by merging the parameters of the component models and adjusting them according to the interpolation coefficients. Despite some technical issues, these approaches are equivalent. In this work, the latter approach was used.

In comparison with the method presented in the next section, the *GMM interpolation* has the advantage that each of the component models λ_k can be estimated only once independently of the target. By properly adjusting the coefficients, the interpolated model can be be quickly adapted to different tasks.

3.2. Data weighting

In the second method proposed, the mixture coefficients are considered during the parameter estimation by maximizing the likelihood function of the interpolated model given by Equation 2. It follows that this problem can be solved by maximizing the so-called auxiliary function $Q(\lambda, \hat{\lambda})$, which is in this case:

$$Q(\lambda, \hat{\lambda}) = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{i=1}^{I} \cdot \gamma_{it}^k \cdot \log(\alpha_k \hat{\omega_i} f_{s_t}(x_t^k|\hat{\mu_i}, \hat{\Sigma_i})) \tag{3}$$

where $\hat{\omega_i}$, $\hat{\mu_i}$ and $\hat{\Sigma_i}$ are, respectively, the mixture coefficient, the mean vector and the covariance matrix of the i-th Gaussian component of the model $\hat{\lambda}$. In this equation, $\gamma_{it}^k = P(i|x_t^k, \lambda)$ is the probability of being in the Gaussian i at time t, given that the model λ generates x_t^k.

The equations used to compute the new parameter estimates $\hat{\lambda} = \{\hat{\omega_i}, \hat{\mu_i}, \hat{\Sigma_i}\}$, can be obtained by taking the respective partial derivatives of $Q(.,.)$. They can be expressed as:

$$\hat{\omega_i} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it}^k \alpha_k}{\sum_{k=1}^{K} N_k \ \alpha_k} \tag{4}$$

$$\hat{\mu_i} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it}^k \alpha_k \ x_t^k}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \ \alpha_k} \tag{5}$$

$$\hat{\Sigma_i} = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it}^k \alpha_k \ (x_t^k - \hat{\mu_i}) \ (x_t^k - \hat{\mu_i})^T}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it}^k \ \alpha_k} \tag{6}$$

The usual ML re-estimation equations can be obtained by setting $K = 1$ [17]. In practice, the presented method was approximated by performing the usual ML estimation with weighted training data sets.

In comparison with the precedent interpolation method, the *data weighting* method has the advantage to generate models with less parameters. On the other hand, it is necessary to re-estimate the model parameters for each different target set.

4. Choice of interpolation coefficients

The interpolation coefficients can be estimated by an approach that is quite similar to the one used on language model interpolation. At the first step, component acoustic models λ_k are estimated for each training subset. Given some held-out data x, an EM algorithm is used to estimate the coefficients, based on the iterative formula [18]:

$$\hat{\alpha_k} = \sum_{t=1}^{T_k} \frac{\alpha_k f_{s_t}(x_t|\lambda_k)}{\sum_{k'=1}^{K} \alpha_{k'} f_{s_t}(x_t|\lambda_{k'})} \tag{7}$$

This task requires an alignment between the data stream and their associated transcriptions in order to calculate the like-

Table 1: *Training corpora information. The size, given in hours, corresponds to the amount of raw data in each subset.*

Subset	Sources	Epoch	Size
train1	RTP channel	2001	64
train2	RTP channel	2010	86
train3	Voice of America	2009	21
train4	Euronews	2009-2010	20
train5	Quaero data	2010	77

Table 2: *Information of the data sets used for evaluation. Last column givens the size of the dev set associated to each of the eval set. Size is given in hours.*

Source	Epoch	Eval		Dev
		Name	Size	Size
RTP channel	2000	eval1	1.3	3.9
Voice of America	2009	eval3	2.0	5.0
Euronews	2010	eval4	2.0	5.0
Quaero Evaluation	2010	eval5	3.5	3.5

lihood $f_{s_t}(x|\lambda_k)$ on each of the component models. This alignment can be performed using either, the manual or the automatic transcriptions of the held-out data. The manner how the automatic transcriptions are generated is a separate issue that is not in the scope of this work. Here, they were obtained using the baseline system with the pooled model.

5. Task and System Overview

The experiments were carried out using the LIMSI speech recognition toolkit, with acoustic, lexical and language models developed for Portuguese and tested with broadcast data.

5.1. Corpora

The audio data used in this work contains about 268 hours of untranscribed data. The relevant information about the corpora is shown in Table 1. The training data were separated into five subsets, according to the source the shows come from and the epoch they were broadcast. For the first experiments, the first 4 sources, shown in the upper part of the table, were grouped into a unique set, henceforth *train1-4*. The *train5* set consists of data collected for the 2011 Quaero Programme Evaluation [1]. Although it consists of data coming from different sources, they were considered to be homogeneous and treated as part of a unique subset. Manual transcriptions of a portion of the *train5* data were available, but they were not used here.

The systems were evaluated on four data sets, coming from the same sources. The relevant information about the evaluation sets is given in Table 2. For each of these sets, an associated development set was used to estimate the interpolation coefficients. The development set of the Quaero data is the only one for which manual transcriptions were available. All the system parameters were tuned on this set. Automatic transcriptions for all the dev sets were generated using the baseline system.

The language model training data include about 640 million words from nine different written sources, such as newspapers, newswires and blogs. These data cover the period from 1991 to 2010. About 30k words of manual transcriptions from RTP shows broadcast in 2000 were also used. The text data was normalized in order to convert numerical forms and abbreviations to spoken forms.

[1] http://www.quaero.org

5.2. System description

The system used in this work is quite similar to other systems developed at LIMSI [19]. It makes use of a pronunciation dictionary, n-gram language models and acoustic models based on continuous density HMMs. Each phone is modeled by a tied-state left-to-right context-dependent triphone HMM, with Gaussian mixture observation densities. Only PLP [20] features were used. The PLP feature vector contains 39 components, including 12 cepstrum coefficients and log energy with their first and second derivatives. The phone set contains 35 phones, as well as special units for silence, breath and hesitation markers.

In this work, the vocabulary, language models and pronunciation dictionary described in [11] were used. The vocabulary was automatically selected based on interpolation of unigram language models. With the selected vocabulary, 2-, 3- and 4-gram component language models were estimated from each of the 10 different sources. The final language models were obtained by interpolation of the component models, with coefficients automatically chosen in order to minimize the perplexity on the Quaero development set. The pronunciation dictionary was obtained via a rule-based grapheme to phoneme converter. For all the experiments, the language models, pronunciation dictionary and decoding parameters were kept fixed. At each test, only the acoustic models were changed.

6. Experiments

The acoustic models were estimated using an unsupervised training approach with the best hypothesis taken as ground truth. The system used to transcribe the training data is described in [10]. In this work, only one iteration of unsupervised training was performed for each model created. All the models trained have the same structure, covering about 15.7k phone contexts and having about 11.5k tied-states. The speech units are modeled by GMMs with up to 32 components, while silence is modeled by mixtures of 1024 components.

6.1. Impact of interpolation coefficients

The first experiments were performed in order to validate the automatic choice of the interpolation coefficients. To better take into account the impact of the coefficients, the training data were separated only into two different sets, *train1-4* and *train5* (See Table 1). Different pairs of coefficients were manually set. Two pairs of coefficients were automatically obtained using the proposed method, with the estimation guided by either, the manual or the automatic transcriptions of the Quaero development set. For each pair of coefficients, a model was estimated using both interpolation methods proposed. The models were evaluated on the *eval5* set.

Table 3 shows the main results obtained. Despite some small differences, the data weighting (DW) and the GMM interpolation (GMMI) methods led to equivalent WER performance levels. The first and last rows show the performances of the component models, trained only on the *train1-4* and *train5* sets. The model estimated on *train5* leads to a WER of 31.7%, which is 2.1% (absolute) smaller than the WER obtained with the model *train1-4*, even if this latter was trained on a subset 2.5 times bigger. Thus, as expected, the *train5* data match better the *eval5* data, since they come from the same source.

Table 3: *WER(%) on the* eval5 *set with acoustic models estimated using different pairs of mixture coefficients defined for the subsets* train1-4 *and* train5*. Data weighting (DW) and GMM interpolation (GMMI) methods were assessed.*

AM	Coefficients		WER(%)	
	train1-4	train5	DW	GMMI
train1-4	1.00	0.00	33.8	
baseline	0.50	0.50	31.7	31.7
auto, 1-best dev	0.22	0.78	**31.3**	31.4
auto, manual dev	0.20	0.80	**31.3**	**31.3**
train5	0.00	1.00	31.7	

The WER obtained with the pooled (baseline) model is given in the second row. The pooled model is equivalent to the weighted model with the coefficients uniformly assigned to the training subsets. This model leads to a WER of 31.7%, which is the same performance obtained with the *train5* model alone. Different WER performance levels were obtained by manually varying the interpolation coefficients, reaching a minimum (31.3%) with the coefficients 0.20/0.80. It turns out that these were the coefficients estimated via the manual transcriptions of the Quaero development data. Both interpolation methods led to the best WER performance in this case. The estimation guided by the automatic transcriptions led to a slight different pair of coefficients (0.22/0.78). However, these coefficients still led to the best performance with the data weighting method. A small absolute loss (0.1%) was obtained with the GMM interpolation method, although the performance obtained (31.4%) is still better than the baseline (31.7%).

These results suggest that the optimal interpolation coefficients can be automatically selected by maximizing the likelihood function on the development data. Furthermore, they can be chosen using an unsupervised approach, in which the estimation is guided by the automatic transcriptions of the dev set.

6.2. Training with all subsets

This section describes the experiments performed to evaluate the two interpolation methods proposed on all of the available test sets. To better take into account the relevance of each of the data sources, the training set was divided into five subsets as indicated in Table 1. For each evaluation set, mixture coefficients were estimated using the automatic transcriptions of the associated development set. The interpolated models were compared with the pooled model and the related MAP adapted models. For each evaluation subset, a MAP model was estimated via the automatic transcriptions of the most similar training subset, with the *prior* distributions obtained from the pooled model.

Table 4 summarizes the results. The second column gives the WERs obtained with the pooled model, which led to an average WER of 22.5% over the four evaluation sets. The last column shows the results obtained with the MAP models. In average, the adapted models led to a slight absolute improvement of 0.1% over the pooled model, but with a gain of performance on two sets, *eval3* (-0.5%) and *eval5* (-0.2%), and a loss on the other two, *eval1* (+0.4%) and *eval4* (+0.1%).

The interpolated models obtained with the automatically estimated coefficients performed better than the equivalent models trained with equally set coefficients for both methods proposed. For all the evaluation sets, the best performance levels were obtained using the data weighting method and the estimated coefficients. Compared to the baseline pooled model, this method led to an average absolute WER reduction of 0.4%, achieving an improvement up to 0.8% on *eval3*. It also led to absolute improvements from 0.2% on *eval4* to 0.6% on *eval1* over the MAP adapted models.

Table 4: *WER(%) on all the evaluation sets obtained with MAP adapted and interpolated models. The interpolated models were obtained using the data weighting (DW) and GMM interpolation (GMMI) methods. In both cases, equally set coefficients (equal) are compared with automatically selected (auto) ones.*

Test set	DW		GMMI		MAP
	equal (pooled)	auto	equal	auto	
eval1	24.7	**24.5**	25.1	25.0	25.1
eval3	14.4	**13.6**	14.3	13.9	13.9
eval4	13.2	**13.1**	13.7	13.6	13.3
eval5	31.7	**31.2**	32.0	31.5	31.5
average	22.5	**22.1**	22.8	22.4	22.4

For all the cases, models estimated using weighted data performed better than the equivalent models interpolated at the decoding phase. A possible explanation to this behavior may be the fact that the parameters of some of the component models were poorly estimated. In particular, two of the component models were trained on only about 20 hours of untranscribed data. This hypothesis can be supported observing the results obtained on *eval5*: when only two subsets were used, no significant difference of performance was observed between the two methods. Nevertheless, for two of the test sets (*eval3* and *eval5*), the models trained with the GMMI method and automatically estimated coefficients performed better than the pooled model. For three sets (except *eval4*), the recognition performance is comparable with the MAP adapted models.

7. Conclusion

This paper has proposed an approach to take into account the relevance of different sources used on acoustic modeling. Two methods to combine subsets of the training data have been presented. In the first one, component models estimated on each source are interpolated at the decoding phase. The second method can be approximated by the parameter estimation with weighted training data. A method to estimate the optimal mixture coefficients using an EM approach has been proposed. It consists in maximizing the likelihood on some acoustic held-out data using either the manual or automatic transcriptions associated to them.

The data weighting method has led to the best WER performances on four different evaluation sets compared to the baseline pooled models, the *maximum a posteriori* adapted models and the GMM interpolated models. Absolute gains of performance up to 0.8% have been observed over the pooled models and up to 0.6% over the MAP adapted models.

This work can be extended in different ways. For instance, instead of estimating a global mixture coefficient for each component model, better performance levels might be achieved by estimating coefficients per phoneme or per phone model. Besides that, the GMM interpolation method could be improved to perform model adaptation at recognition time. It could be done by assigning mixture coefficients to each show (or speaker) after a first decoding pass, at the condition that each of the component models are well-trained.

8. References

[1] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains," *AT&T Technical Journal*, vol. 64, no. 6, July-August 1985.

[2] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition," in *ICASSP*, Hong Kong, April 2003, pp. I–212–215.

[3] V. Fischer, E. Janke, S. Kunzmann, and T. Ross, "Multilingual acoustic models for the recognition of non-native speech," in *Automatic Speech Recognition and Understanding (ASRU). IEEE Workshop on*. IEEE, 2001, pp. 331–334.

[4] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *ICASSP*, vol. 1. IEEE, 2003, pp. I–540.

[5] Y. Oh, J. Yoon, and H. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," *Speech Communication*, vol. 49, no. 1, pp. 59–70, 2007.

[6] D. Vergyri, L. Lamel, and J.-L. Gauvain, "Automatic Speech Recognition of Multiple Accented English Data," in *InterSpeech'10, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, 2010.

[7] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[8] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[9] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Maximum a posteriori adaptation for large scale hmm recognizers," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 725–728.

[10] T. Fraga-Silva, L. Lamel, and J.-L. Gauvain, "Lattice-based unsupervised acoustic model training," in *ICASSP*, Prague, Czech Republic, May 2011, pp. 4656–4659.

[11] T. Fraga-Silva, V.-B. Le, L. Lamel, and J.-L. Gauvain, "Incorporating MLP features in the unsupervised training," in *SLTU*, Cape Town, South Africa, May 2012.

[12] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *Speech Prosody 2004, International Conference*, 2004.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in hmm-based speech synthesis system," in *Proc. Eurospeech*, vol. 97. Rhodes, Greece: ESCA, 1997.

[14] T. Tan and L. Besacier, "Acoustic model interpolation for non-native speech recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1009.

[15] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, February 1998, pp. 301–305.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[17] S. Young and G. Bloothooft, *Corpus-based methods in language and speech processing*. Kluwer Academic Publishing, 1997, vol. 2.

[18] R. DeMori and M. Federico, "Language model adaptation," *NATO ASI series. Series F: Computer and System Sciences*, pp. 280–303, 1999.

[19] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[20] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.