JOINT ANALYSIS OF VOCAL TRACT LENGTH AND TEMPORAL INFORMATION FOR ROBUST SPEECH RECOGNITION

*Chien-Lin Huang, Chiori Hori, Hideki Kashioka, Bin Ma**

National Institute of Information and Communications Technology, Kyoto, Japan
*Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
{chien-lin.huang, chiori.hori, hideki.kashioka}@nict.go.jp, mabin@i2r.a-star.edu.sg

ABSTRACT

This paper presents a joint analysis approach to address the acoustic feature normalization for robust speech recognition. The variations in acoustic environments and speakers are the major challenge for speech recognition. The conventional normalizations of these two variations are separately processed, applying the speaker normalization with an assumption of a noise free condition and applying the noise compensation with an assumption of speaker independency, and thus resulting in a suboptimal performance. The proposed joint analysis approach simultaneously considers the vocal tract length normalization and averaged temporal information of cepstral features. In a data-driven manner, the Gaussian mixture model is used to estimate the conditional parameters in the joint analysis. Experimental results show that the proposed approach achieves a substantial improvement.

Index Terms— Joint analysis, vocal tract length normalization, speech recognition, feature normalization

1. INTRODUCTION

Variations of acoustic environments and speakers are the major challenge in current speech recognition systems [1]–[4]. Speech recognition systems may work well under a clean acoustic environment but their performance degrades dramatically in adverse acoustic conditions. To achieve the robust speech recognition, many techniques have been proposed to address the acoustic mismatch problem in both model and feature space while we focus on the feature space in this paper. Cepstral mean normalization was used to remove the global shift of cepstral features and compensate for the main effect of channel distortions [5] and cepstral variance normalization was used to compensate the linear channel variations in feature analysis [6]. The histogram equalization (HEQ) provides a transformation mapping the histogram of each feature component onto a reference histogram to compensate the noise effect [7]. In order to recover the clean speech features, RASTA [8, 9], multiple microphones [10] and Kalman filters [11] have common been employed for noise robust speech recognition. Among the techniques, a simple auto-regression moving-average (ARMA) filtering has demonstrated its effectiveness for noise reduction by averaging the temporal information [12]–[14]. Besides the acoustic environment variability, the speaker variability also affects the speech recognition performance much. To address this problem, the vocal tract length normalization (VTLN) algorithm [15] normalizes the difference of vocal tract traits between speakers so that the extracted acoustic features are robust to variations in vocal tract length, and thus the performance of speaker independent speech recognition is improved by accounting for inter-speaker variability [16]–[18].

In general, the VTLN and ARMA are separately processed to compensate the speaker variability and noisy condition in speech recognition. The acoustic condition is assumed to be noise free when VTLN is applied, while speaker mismatches are ignored when ARMA is conducted. Obviously it is not an optimal way for the estimation. In this paper, we propose a joint analysis in feature normalization where the vocal tract length normalization and averaged temporal information of Mel-frequency cepstral coefficients are considered at the same time. The effectiveness of this joint analysis in feature normalization has been verified in experiments using AURORA 2. In the following, we present the proposed joint analysis of VTLN and ARMA in Section 2. Section 3 shows experiments in detail. We conclude with a summary of findings in Section 4.

2. JOINT ANALYSIS OF VTLN AND ARMA

We investigate a feature extraction method of a joint analysis of VTLN and ARMA temporal filtering for speech recognition as shown in Fig. 1. Mel-frequency cepstral coefficients (MFCCs) are adopted as acoustic features [19]. Fast Fourier transform (FFT) is firstly applied to estimate the speech spectra, followed by vocal tract length normalization, Mel-cepstral analysis and ARMA temporal filtering. Based on the Gaussian mixture model (GMM), the warping factor of VTLN and the order of ARMA temporal filtering are estimated by the maximum likelihood criterion. The main contribution of this joint analysis is to normalize

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

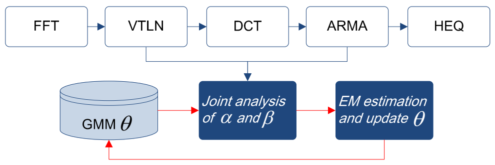
FFT
VTLN
DCT
ARMA
HEQ
GMM θ
Joint analysis of α and β
EM estimation and update θ

Fig. 1. Joint analysis of VTLN and ARMA.

speaker and noise factors at the same time. The joint analysis is estimated based on the expectation-maximization (EM) training algorithm. Finally, acoustic features are normalized to zero mean and unity variance using HEQ for the higher discrimination ability. The target distribution of HEQ is selected as a Gaussian.

2.1. VTLN on the Speaker-Specific Mel Scale

In MFCC feature extraction, the frequency bins are smoothed with the perceptually motivated Mel-frequency scaling after the log-amplitude of the magnitude spectrum. To normalize different vocal tract lengths between speakers, the VTLN based on the speaker-specific Mel scale is estimated as follows:

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700\times\alpha}\right) \tag{1}$$

where the warping factor α is used to adjust a speaker-specific Mel scale. This frequency-warping procedure is implemented as a filter bank modification. $\alpha > 1.0$ results in a compressed spectrum, $\alpha < 1.0$ results in a stretched spectrum, and $\alpha = 1.0$ is for a non-warped spectrum. Note that speech signals of female speakers tend to have shorter vocal tract lengths and higher formant frequencies than male speakers [17]. One would expect to see more compressed spectra in female speech than in male speech. The warping factor that maximizes the likelihood of the speech utterance is searched in the maximum likelihood (ML) manner, i.e., a HMM decoding is required. To reduce the computational cost, the HMM decoding can be replaced by a broad-class GMM decoding. In this paper, we further simplify the VTLN process in the proposed joint analysis method with the GMM and the whole utterance is used to estimate the warping factor. With the GMM θ, the computational cost of VTLN is reduced much.

2.2. Auto-regressive Moving Average Filtering

To reduce the noise effect, a non-causal auto-regressive moving average filter is used in cepstral feature extraction as the moving averages of temporal information defined as:

$$\tilde{\mathbf{c}}_t = \begin{cases} \dfrac{\sum_{i=-\beta}^{\beta}\mathbf{c}_{t+i}}{2\beta+1} & if\ \beta < t \le T-\beta \\ \mathbf{c}_t & otherwise \end{cases} \tag{2}$$

where β is the order of the ARMA filter and $\mathbf{c}$ are cepstral coefficients. The ARMA filter is a low-pass filter, smoothing out any spikes in the time sequence. A small β retains the short-time cepstral information but is more vulnerable to noises, while a large β makes the processed features less corrupted by noises but at the cost of losing some short-time cepstral information. There is an inherent trade-off to decide the order of the ARMA filter. The order of the ARMA filter is empirically selected to fit a corpus or an evaluation condition with a fixed value [12]. Unlike fixed values of the ARMA estimation, we propose an adaptive method to dynamically decide the order β based on a joint analysis with the EM estimation.

2.3. Joint Analysis with the EM Estimation

In the proposed joint analysis of VTLN and ARMA, the warping factor α of VTLN and the order β of the ARMA filter are estimated in a data-driven manner by maximizing the likelihood of observation feature vectors $\mathrm{x}^{\alpha,\beta}$, given the GMM θ,

$$p(\mathrm{x}^{\alpha,\beta} \mid \theta) = \sum_{m=1}^{M} w_m\, p(\mathrm{x}^{\alpha,\beta} \mid \lambda_m) \tag{3}$$

where $\mathrm{x}^{\alpha,\beta} = [x_1,..,x_d]^T$ is the d-dimensional feature vector and the weight of Gaussian component $\sum_{m=1}^{M} w_m = 1$ while M is set to 64 in this study. The mixture model $p(\mathrm{x}^{\alpha,\beta} \mid \lambda_m)$ is a normal distribution, with each Gaussian component represented by the parameters $\lambda_m = \{\mu_m, \Sigma_m\}$, where μ_m and Σ_m are the mean vector and covariance matrix. We view the joint analysis as two estimations: One is the joint analysis of parameters α and β. The other is the EM estimation for updating the GMM parameters, θ. The joint analysis of parameters α and β is performed for each utterance shown in Algorithm 1. The optimized α and β are then used to estimate speaker and noise normalized speech feature and update the GMM, $\theta = \{w_1,..,w_M,\ \lambda_1,..,\lambda_M\}$ by an iterative EM algorithm. The EM training algorithm consists of an expectation step (**E-step**) and a maximization step (**M-step**) which are iteratively estimated until the equation

$$\log p(\mathrm{X} \mid \theta) = \log \prod_{n=1}^{N} p(\mathrm{x}_n^{\alpha,\beta} \mid \theta) \tag{4}$$

converges to an optimum, where $\mathrm{X} = \{\mathrm{x}_1^{\alpha,\beta},..,\mathrm{x}_N^{\alpha,\beta}\}$ denote N i.i.d. samples. The K-means algorithm has been applied to initialize the parameters $\theta^{t=0}$ based on $\alpha = 1$ and $\beta = 0$.

E-step: The data X are assumed to be incomplete and the complete data $\varphi = (\mathrm{X}, \mathrm{Z})$ are determined by estimating the set of variables $\mathrm{Z} = \{\mathrm{z}_1,..,\mathrm{z}_M\}$ where z_m is the M-

Algorithm 1: Joint analysis of parameters α and β

Step 0. Set the iteration index $t=0$ and determine the optimization function

$$Q(\alpha,\beta)=p(\mathrm{x}^{\alpha,\beta}\mid\theta)$$

- Set initial values of parameters $(\alpha,\beta)=(1,0)$
- Parameters α and β are independent

Step 1. Fix β^*, optimize for α,

$$\alpha_{t+1}=\alpha_t+\lambda_\alpha\left.\frac{\partial Q(\alpha,\beta^*)}{\partial\alpha}\right|_{\alpha=\alpha_t}$$

In real application, $\partial Q(\alpha,\beta^*)/\partial\alpha\overset{\Delta}{=}Q(\alpha+\Delta\alpha,\beta^*)-Q(\alpha,\beta^*)$

Step 2. Fix α^*, optimize for β,

$$\beta_{t+1}=\beta_t+\lambda_\beta\left.\frac{\partial Q(\alpha^*,\beta)}{\partial\beta}\right|_{\beta=\beta_t}$$

In real application, $\partial Q(\alpha^*,\beta)/\partial\beta\overset{\Delta}{=}Q(\beta+\Delta\beta,\alpha^*)-Q(\beta,\alpha^*)$

Step 3. Set $t=t+1$, and Go to Step 1.

dimensional probability vector. The log likelihoods of complete data φ are estimated by

$$\log p(\varphi\mid\theta)=\sum_{n=1}^{N}\sum_{m=1}^{M}\mathrm{z}_m^n\log[w_m p(\mathrm{x}_n^{\alpha,\beta}\mid\lambda_m)] \tag{5}$$

where $\mathrm{z}_m^n=P(m\mid\mathrm{x}_n^{\alpha,\beta},\theta^t)=\dfrac{w_m^t p(\mathrm{x}_n^{\alpha,\beta}\mid\lambda_m^t)}{\sum_{i=1}^{M}w_i^t p(\mathrm{x}_n^{\alpha,\beta}\mid\lambda_i^t)}$ is the posterior probability estimated after *t* iterations.

M-step: The parameters θ^{t+1} are estimated based on the variables z_m^n. We obtain

$$\begin{aligned}&w_m^{t+1}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{z}_m^n,\quad \mu_m^{t+1}=\frac{\sum_{n=1}^{N}\mathrm{z}_m^n\mathrm{x}_n^{\alpha,\beta}}{\sum_{n=1}^{N}\mathrm{z}_m^n},\\ &\text{and}\ \ \Sigma_m^{t+1}=\frac{\sum_{n=1}^{N}\mathrm{z}_m^n(\mathrm{x}_n^{\alpha,\beta}-\mu_m^{t+1})(\mathrm{x}_n^{\alpha,\beta}-\mu_m^{t+1})^T}{\sum_{n=1}^{N}\mathrm{z}_m^n}\end{aligned} \tag{6}$$

Instead of a consecutive optimization by finding optimal α for a fixed β, subsequently optimizing β for a fixed α, we found in experiments that an exhaustive search for both parameters that maximize the likelihood can reduce the computational cost and give a similar performance. The search can be made in the range $0\le\beta\le8$ with a step size of 1 for ARMA filtering, and the range $0.6\le\alpha\le1.24$ with a step size of 0.04 for VTLN. Acoustic features are generalized across a variety of noise and speaker variations based on the joint analysis. The advantage of this joint analysis approach is that it is entirely data driven. Unlike the fixed order of the ARMA filter or individual estimation of α and β, the joint analysis is an adaptive method to dynamically optimize parameters based on the ML estimation.

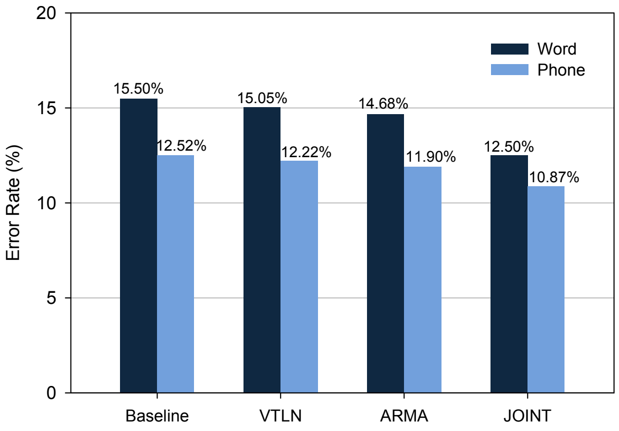

Fig. 2. The proposed joint analysis approach (JOINT) reduces the averaged word and phone error rates to 12.50% and 10.87%, respectively, with a multi-condition training and over all testing data (N1~N8).

3. EXPERIMENTS

We applied the joint analysis to AURORA 2 database [20]. Five signal-to-noise (SNR) conditions were evaluated including 5dB, 10dB, 15dB, 20dB and clean. For each SNR level, there are eight types of noise extracted from Set A and Set B of AURORA 2 database including subway, babble, car, exhibition, restaurant, street, airport and train-station. They are represented as abbreviations N1, N2,..., N8. Experiments are reported with the word or phone error rate considering insertion, deletion and substitution errors [21]. The speech data were decoded via HMM acoustic models trained with clean and multi condition data, respectively. Each HMM contained five states and the number of Gaussian mixture components per state ranged from 2 to 32 based on the quantity of the training data. Each frame of the speech data is represented by a 36-dimensional feature vector, consisting of 12 MFCCs, together with their deltas and double-deltas. HEQ was applied after the MFCC extraction as the baseline. VTLN, ARMA and the proposed joint analysis approach (JOINT) were further added to the baseline system. For a fair comparison, each of VTLN and the ARMA filter was optimized using the same technique as the joint optimization proposed in the paper.

3.1. Effects of the Joint Analysis

Fig. 2 shows the performance in multi-condition training with all testing data. The averaged word error rate of the

Table 1. Word recognition accuracy (%) of the baseline, VTLN, ARMA, and the proposed joint analysis (JOINT) in different SNR levels and with noisy types

SNR	Approach	Noise Type								
		N1	N2	N3	N4	N5	N6	N7	N8	avg.
5dB	Baseline	75.30	74.70	76.60	76.70	73.80	75.70	76.70	75.70	75.65
	VTLN	71.40	75.70	77.80	76.50	73.50	76.40	77.90	75.70	75.61
	ARMA	80.60	75.60	77.40	77.30	74.20	75.90	77.30	76.40	76.84
	JOINT	82.90	81.70	83.00	79.80	73.80	76.20	79.80	77.30	79.31
10dB	Baseline	83.60	82.30	83.10	83.60	81.30	82.80	82.50	82.90	82.76
	VTLN	84.00	83.40	83.70	84.50	83.00	82.70	83.60	83.70	83.58
	ARMA	85.00	83.80	84.70	84.60	81.30	84.30	84.30	84.40	84.05
	JOINT	86.10	88.10	87.10	86.90	82.50	82.70	84.60	82.70	85.09
15dB	Baseline	86.10	85.60	85.60	86.80	85.80	86.20	86.30	87.20	86.20
	VTLN	86.60	87.10	86.20	88.10	87.70	86.30	86.90	86.90	86.98
	ARMA	87.40	86.70	87.90	87.30	87.20	86.50	87.10	88.30	87.30
	JOINT	88.60	89.40	89.20	89.60	86.50	85.30	86.10	87.80	87.81
20dB	Baseline	87.70	88.00	88.40	87.50	87.50	88.70	87.10	88.60	87.94
	VTLN	89.20	89.40	88.10	89.00	89.30	88.90	88.10	87.90	88.74
	ARMA	89.40	89.50	88.60	88.50	89.40	89.30	89.50	89.90	89.26
	JOINT	89.30	91.80	90.10	90.80	88.50	87.40	89.30	88.00	89.40
clean	Baseline	87.70	88.50	87.20	89.00	87.70	88.50	87.20	89.00	88.10
	VTLN	88.40	89.30	88.10	89.10	88.40	89.30	88.10	89.10	88.73
	ARMA	89.00	89.80	88.80	89.60	89.00	89.80	88.80	89.60	89.30
	JOINT	91.40	91.90	91.90	92.50	88.10	87.50	88.20	87.60	89.89

baseline system (with HEQ) is 15.50% under various noise conditions. Compared with the baseline system, the systems with VTLN and ARMA achieve relative word error rate reductions of 2.91% and 5.31%, respectively. The ARMA normalization slightly outperforms the VTLN normalization in average. With the proposed joint analysis, the overall relative word error rate reduction 19.35% (from 15.50% to 12.50%) is achieved. The experimental results demonstrate that joint analysis approach is a more robust feature normalization approach against the mismatch of acoustic environments and the difference of vocal tract lengths of speakers.

3.2. Different Noisy Environments and Techniques

To investigate the performance under various testing conditions in detail, Table 1 summarizes the word accuracy of speech recognition systems based on the baseline, VTLN, ARMA and the joint analysis (JOINT) under different SNR levels, 5dB, 10dB, 15dB, 20dB and clean. The results were based on the multi-condition training data. It shows that the joint analysis outperforms individual VTLN or ARMA in most testing conditions. It is also noted that the joint analysis achieves bigger improvements for test Set A (N1~N4) than Set B (N5~N8). Since the GMM θ was estimated with the training data containing the same four types of noises as that of test Set A, it is reasonable that the joint analysis gives bigger improvements for test Set A than test Set B. Identifying a way to enhance the generalization capability of the joint analysis in testing conditions with unseen types of noises is an interesting task to be further explored in the future.

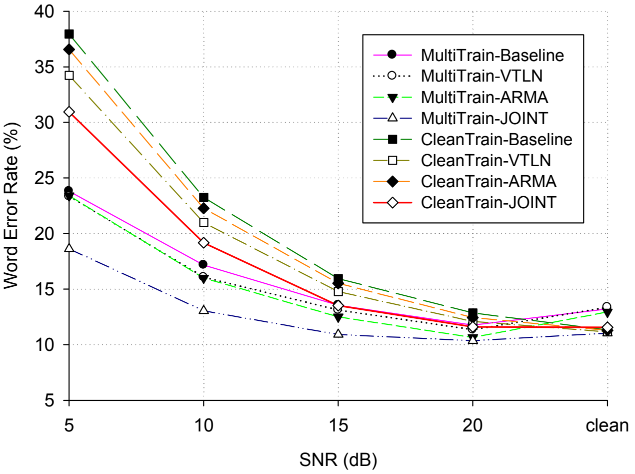

Fig. 3. Word error rates (%) of different SNRs using multi- and clean- condition training data.

3.3. Multi-Condition and Clean-Condition Training

We made the comparison using multi-condition training and clean-condition training (MultiTrain & CleanTrain) data as shown in Fig. 3, where the results were the average over the test data with different noises. We take into consideration of noisy environments in acoustic model training. Several observations can be found. The joint analysis shows a better performance while the performance gap between the joint analysis and other approaches is much more noticeable under the low SNR conditions. The multi-condition training generally outperforms the clean-condition training under the low SNR conditions. In addition, the proposed joint analysis approach has achieved the lowest word error rates under both multiple and clean training conditions.

4. CONCLUSION

We have investigated a robust feature normalization method for speech recognition based on the joint analysis of vocal tract length normalization and averaged temporal information of spectral features. To alleviate mismatches of speakers and noise environments and to avoid suboptimal parameter estimations with separated VTLN and ARMA processes, the joint analysis approximates the bias between clean and noisy speech and the different vocal tract lengths of speakers based on the GMM which is estimated by the EM training algorithm under the ML criterion. Experimental results confirm that the proposed joint analysis approach can give an obvious performance improvement. The averaged relative word error rate reduction over the baseline is 19.35% under various training and testing conditions.

5. REFERENCES

[1] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, “Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks,” in *Proc. ICLR*, 2013.

[2] C. Kim and R. M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech recognition,” in *Proc. ICASSP*, 2012.

[3] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, “Speaker Clustering Using Vector Representation with Long-Term Feature for Lecture Speech Recognition,” in *Proc. ICASSP*, 2013.

[4] X. Xiao, E. S. Chng, and H. Li, “Joint Spectral and Temporal Normalization of Features for Robust Recognition of Noisy and Reverberated Speech,” in *Proc. ICASSP*, 2012.

[5] B. S. Atal, “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal Acout. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[6] O. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, vol. 25, pp. 133–147, 1998.

[7] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez, and A. J. Rubio, “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[8] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] S. Vuuren and H. Hermansky, “Data-driven design of RASTA-like filters,” in *Proc. Eurospeech*, 1997.

[10] T. M. Sullivan, “Multi-microphone Correlation-based Processing for Robust Automatic Speech Recognition,” Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1996.

[11] K. K. Paliwal and A. Basu, “A Speech Enhancement Method based on Kalman Filtering,” in *Proc. ICASSP*, pp. 177–180, 1987.

[12] C.-P. Chen and J. Bilmes, “MVA Processing of Speech Features,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[13] C.-P. Chen, J. Bilmes, and D. Ellis, “Speech Feature Smoothing for Robust ASR,” in *Proc. ICASSP*, 2005.

[14] C.-P. Chen, J. Bilmes, and K. Kirchhoff, “Low-Resource Noise-Robust Feature Post-Processing on Aurora 2.0,” in *Proc. ICSLP*, 2002.

[15] T. Kamm, G. Andreou, and J. Cohen, “Vocal tract normalization in speech recognition: Compensating for systematic speaker variability,” in *Proc. of the 15th Annual Speech Research Symposium*, 1995.

[16] P. Zhan and A. Waibel, “Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition,” *Technical Report: CMU-LTI-97-150*, 1997.

[17] L. Lee and R. Rose, “A Frequency Warping Approach to Speaker Normalization,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[18] G. Saon and J.-T. Chien, “Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012.

[19] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Englewood Cliffs, N.J.: Prentice Hall, 2001.

[20] N. Parihar and J. Picone, “Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02,” *Institute for Signal and Information Processing, Mississippi State Univ., Mississippi, MS, Technical Report,* 2002.

[21] C.-L. Huang and C.-H. Wu, “Generation of Phonetic Units for Mixed-language Speech Recognition based on Acoustic and Contextual Analysis,” *IEEE Trans. Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.