Interspeech 2018
2-6 September 2018, Hyderabad

Robust Mizo Continuous Speech Recognition

Abhishek Dey[1], *Biswajit Dev Sarma*[5], *Wendy Lalhminghlui*[3], *Lalnunsiami Ngente*[3], *Parismita Gogoi*[2,3], *Priyankoo Sarmah*[3], *S.R.M. Prasanna*[3,4], *Rohit Sinha*[3] *and S.R. Nirmala*[1]

[1]GUIST, Gauhati University, Guwahati-781014, India
[2]DUIET, Dibrugarh University, Dibrugarh-786004, India
[3]Indian Institute of Technology Guwahati, Guwahati-781039, India
[4]Indian Institute of Technology Dharwad, Dharwad-580011, India
[5]Kaliber Labs Inc, Guwahati-781039, India

{abhishekdey.gu,biswajit.devsarma,nirmalasr3}@gmail.com,
{wendy,siami,parismitagogoi,priyankoo,rsinha}@iitg.ac.in, prasanna@iitdh.ac.in

Abstract

Mizo is an under-resourced tonal language that is mainly spoken in North-East India. It has 4 canonical tones along with a tone-sandhi. In Mizo language, a majority of the words contain tone information. As a result of that, it exhibits higher acoustic variability like other tonal languages in the world. In this work, we investigate the impact of tonal information on robust Mizo continuous speech recognition (CSR). First, separate baseline CSR systems are developed employing the Mel-frequency cepstral coefficient (MFCC) based acoustic features and salient acoustic modeling paradigms. For further improvement, the tonal information has been incorporated in each of the CSR systems. For this purpose, 3-dimensional tonal features are derived which include pitch, pitch-difference, and probability of voicing values. Our experimental study reveals that with the inclusion of tonal information, the robustness of Mizo CSR system gets enhanced across all acoustic modeling paradigms. This trend is attributed to lesser degradation in the fundamental frequency information than the vocal tract information under noisy conditions.

Index Terms: Mizo language, tonal information, robust CSR

1. Introduction

Continuous speech recognition (CSR) for tonal languages have been extensively studied in languages like Cantonese, Mandarin, Thai, Vietnamese. Previous studies have reported that tones carry important lexical information in tonal languages which serves as an important cue for speech recognition tasks [1–7]. In Cantonese speech recognition task, 8% relative improvement in character error rate is achieved using weighted tone information [2]. Cao *et al.* reported that tone information integration helped in achieving a 16.2% relative character error rate reduction in continuous Mandarin speech recognition [4]. Similarly, in Mandarin Broadcast News speech recognition task, Lei *et al.* [5] reported the reduction of the character error rate from 13.0% to 11.5% on CTV test set. In case of Thai spelling recognition task, it has been reported that inclusion of tone information resulted in 23.85% error rate reduction from the baseline system [6]. Similarly 28.6% relative reduction in word error rate is reported when tone information is incorporated in Vietnamese continuous speech recognition system [7]. In case of noisy speech conditions, the vocal tract information gets severely affected, yet the fundamental frequency (F0) information is largely retained. Drugman *et al.* [8] reported that excitation source based features are robust to noisy conditions. Yasui *et al.* [9] achieved the improvement in word accuracy from 43.2% to 52.5% with inclusion of F0-based tonal features when evaluated on noisy JNAS database test set.

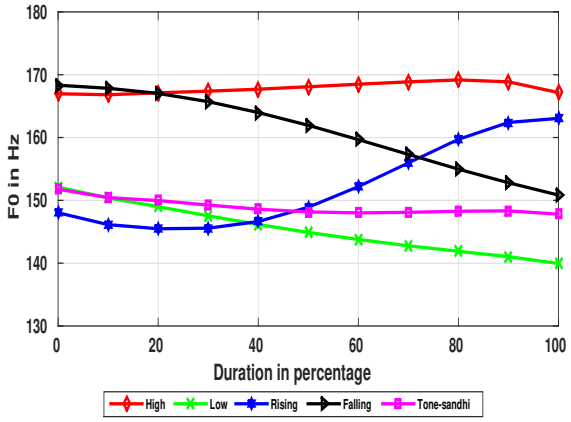

Figure 1: *F0 contours of four contrastive tones with tone-sandhi in Mizo. The data involved in this tonal analysis is described in [10, 11].*

In this work, we have considered Mizo as the language of study and explored the effectiveness of tone information in the context of Mizo continuous speech recognition task. Apart from using tone related features, we have also explored modeling the acoustic-phonetic units using tone information. Mizo is an ethnolinguistic term which stands for both the language and the tribe. Genetically, Mizo belongs to the Kuki-Chin-Naga subgroup of the Tibeto-Burman language family spoken by about 700,000 speakers in the state of Mizoram and its neighbouring states in North East of India. Apart from India, it is also spoken in neighbouring countries like Myanmar and Bangladesh. Preliminary studies reported that Mizo has four canonical tones namely high, low, rising and falling [12–15]. As shown in Figure 1, the four tones in Mizo are found to be significantly different from each other in terms of their F0 slopes [14]. Hence, along with the F0 contour, the F0 slope plays a significant role incorrect identification of tones in Mizo [16]. Accordingly, a method for automatic detection of tones was proposed, resulting in about 70% accuracy incorrect classification of the four tones in Mizo [17]. Apart from the four canonical tones, there is tone-

10.21437/Interspeech.2018-2125

Table 1: *Description of Mizo speech corpus used in this study*

Entity	Training	Testing
Duration of speech data (in Hrs)	5.5	1.5
No. of utterances	7394	1568
No. of speakers	62	19
No. of male speakers	31	12
No. of female speakers	31	7

sandhi in Mizo where the presence of high and falling tone after a rising tone lowers the F0 of the rising tone [11, 12, 15, 18].

The remainder of the paper is organized as follows. Section 2 describes the speech corpus used in this work. Section 3 discusses the tones in Mizo. Section 4 discusses the experimental set up. Section 5 discusses the experimental results and finally, the paper is summarized and concluded in Section 6.

2. Speech Corpus

The speech corpus is collected from 81 native Mizo speakers covering a vocabulary size of 670 words. The speech data is recorded in a sound treated booth with a Shure SM10A head-mounted, close talk microphone connected to a Tascam DR100-MKII recorder. The elicited production is digitally recorded at a sampling frequency of 16 kHz and bit resolution of 16 bits/sample. The recorded speech files are chunked into small segments of $3 - 5$ second duration, depending on the pauses. The reading material used in this corpus consists of 9 different Mizo passages. Each passage consists an average of 38 sentences. Each speaker reads a minimum of two passages. The entire speech corpus is divided into two parts in approximately $8 : 2$ proportion. The major portion of the speech corpus is used for training the acoustic models while the minor portion is used for evaluating the efficacy of the trained phone models. While dividing the speech corpus into two parts, it is ensured that the speakers in the training set are different from those in the testing set. The descriptive statistics of the speech corpus used in this study is provided in Table 1.

3. Tones in Mizo

The tone bearing unit in Mizo is the rhyme of a syllable where the nucleus is vowel(s) or vowel(s) nucleus and sonorant coda. As mentioned earlier, Mizo has four different tones, namely, high, low, rising, and falling with a rising tone-sandhi. The F0 contours of the four canonical tones and sandhi tone in Mizo are shown in Figure 1. As seen in Figure 1, high tone and tone-sandhi have static F0 contours while low, falling and rising tone show dynamic F0 contours. The F0 contours of low tone and tone-sandhi start at the same point and the low tone falls gradually from about 20% of the total duration till the end while tone-sandhi remains level. At the same time, the initiation point of rising tone is close to the initiation points of low tone and tone-sandhi, with a short downward dip initially which later starts to rise from about 30% until 90% of the total duration. Falling and high tones, on the other hand, share the same starting point, falling tone shows a gradual falling F0 contour pattern throughout the total duration while high tone shows a pattern of level F0 contour. The four canonical Mizo tones: high, low, rising and falling along with tone-sandhi are indicated as H, L, R, F, and S, respectively throughout the paper. There are 36 phonetic units defined in Mizo. The complete phonetic inventory along with the tonal annotation is shown in Table 2.

Table 2: *Inventory of Mizo phonetic units*

Phonetic units in IPA	Phonetic units in ASCII	Phonetic units with their corresponding tones in ASCII
a	a	a, a-F, a-H, a-L, a-R
b	b	b
ɔ	c	c, c-F, c-H, c-L, c-R
d	d	d
e	e	e, e-F, e-H, e-L, e-R
f	f	f
h	h	h
l̥	hl	hl
m̥	hm	hm
n̥	hn	hn
ŋ̊	hng	hng
r̥	hr	hr
i	i	i, i-F, i-H, i-L, i-R, i-S
k	k	k
kʰ	kh	kh
l	l	l, l-F, l-H, l-L, l-R
m	m	m, m-F, m-H, m-L, m-R, m-S
n	n	n, n-F, n-H, n-L, n-R, n-S
ŋ	ng	ng, ng-F, ng-H, ng-L, ng-R, ng-S
o	o	o-F, o-H, o-L, o-R, o-S
p	p	p
pʰ	ph	ph
r	r	r, r-F, r-H, r-L, r-R, r-S
s	s	s
t	t	t
tʰ	th	th
tl	tl	tl
tl̥ʰ	tlh	tlh
tr	tr	tr
tr̥ʰ	trh	trh
ts	ts	ts
tsʰ	tsh	tsh
u	u	u-F, u-H, u-L, u-R
v	v	v
ʔ	x	x
z	z	z

4. Experimental Setup

In this section, we discuss the different acoustic modeling paradigms that have been adopted in this study. We have explored the deep neural network (DNN) and subspace Gaussian mixture model (SGMM) based acoustic modeling approaches in addition to the Gaussian mixture model (GMM) based approach. All the experimental evaluations are conducted using the Kaldi speech recognition toolkit [19].

4.1. Front-end features

Experimental evaluations are carried out using one of the most widely used front-end, the Mel frequency cepstral coefficients (MFCC) [20]. The MFCC feature vectors are computed from Hamming windowed frames of 25 ms duration with a frame shift of 10 ms and pre-emphasis factor of 0.97. Each feature vector consists of log-energy and 12 MFCCs ($C1$-$C12$). In order to capture the dynamic characteristics of the vocal tract system, 13-dimensional static MFCC features are appended with their velocity and acceleration components. The resulting 39-dimensional final vector is used in the acoustic modeling.

4.2. GMM-HMM system

This system is initialized with a context-independent monophone acoustic model using 39 dimensional MFCC feature vectors. Each of the 36 phonetic units listed in the second column of Table 2 is modeled by a 3 state left-to-right HMM model where the probability density function of each state is a GMM with 8 mixtures. In order to capture contextual information, cross-word triphone acoustic models are trained with decision tree-based state tying. The 13 dimensional static MFCC feature vectors are spliced in time in order to capture the dynamic information implicitly from feature vectors of the neighbouring frames. Four frames to the left and 4 frames to the right of the central frame are spliced thereby making the total feature dimension to 117 (13×9). The dimension of the spliced feature vector is then reduced from 117 to 40 through LDA. The resulting feature vectors are further decorrelated using MLLT. The derived feature vectors are normalized using cepstral mean and variance normalization. The extracted features are further normalized using fMLLR [21] and speaker adaptive training (anastasakos1997speaker) [22] is employed using fMLLR transformations.

4.3. SGMM-HMM system

In conventional GMM-HMM systems, a large number of model parameters are required to be estimated. This problem is well addressed by the SGMM [23] based acoustic modeling framework. In this approach, the complex distribution of parameters is represented in a compact way. Here, the HMM states globally share a common structure, and only the state-dependent parameters are required to be estimated. Instead of estimating GMM parameters directly from the training data, the model parameters are derived from the low-dimensional model and speaker subspaces that can capture phonetic and speaker correlations. As a result of this, the total number of parameter estimation is reduced, which makes it possible to learn the model parameters with a limited amount of training data. Here the unit distributions are derived from a universal background model (UBM). In the current work. 400 Gaussians are selected for training the UBM.

4.4. DNN-HMM system

We also explored the DNN-HMM-based acoustic modeling approach in this study. A feed-forward deep neural network is trained using multiple hidden layers that takes time-spliced feature vectors with LDA+MLLT+fMLLR as input and computes the posterior probabilities over HMM states as output. The specification of the parameters used in training the DNN-HMM system is detailed in Table 3.

4.5. Tone features and tone modeling

Since Mizo is a tonal language, we have also extracted 3-dimensional tonal features i.e., pitch, pitch-difference, and probability of voicing (POV). The tonal features are extracted using Kaldi pitch tracker [24]. It is a highly modified version of the robust automatic pitch tracking algorithm (RAPT) that assigns a pitch not only to voiced frames but also to unvoiced frames while constraining the pitch trajectory to be continuous. The algorithm produces a quantity that can be used as a probability of voicing measure that is based on finding lag values that maximize the normalized cross-correlation function (NCCF). The 3-dimensional tonal features are appended to the 13-dimensional MFCC features. The resulting

Table 3: *Parameter specifications of DNN-HMM system*

Parameter	Specification
No. of hidden layers	3
No. of epochs	20
Dimension of hidden layer	1024
Mini batch size	128
Initial learning rate	0.01
Final learning rate	0.0015

Table 4: *Phone level break up of an example word in the lexicon with and without tone information*

Lexicon	Word	Phone level break up
Without tone information	Mizoram	m i z o r a m
With tone information	Mizoram	m i-L z o-H r a m-H

16-dimensional base feature vectors are time spliced with a context of 9 frames considering 4 frames to the left and 4 frames to the right of the central frame. Using linear discriminant analysis (LDA) [25, 26], the dimension of the spliced feature vectors are reduced from 144 (16×9) to 40. The resulting feature vectors are further decorrelated using maximum likelihood linear transform (MLLT) [27, 28] and feature-space maximum likelihood linear regression (fMLLR) [21].

The baseline CSR system is modeled using 36 phonetic units as listed in the second column of Table 2. However, these phonetic units do not model the tone information. In order to capture the tone information in the language, the phonetic units are modeled along with their associated tone information as listed in the third column of Table 2. For illustrating the implementation of tone modeling, an example word in the lexicon with and without tone information is shown in Table 4.

5. Results and discussion

For evaluating the noise robustness of Mizo speech recognition system, three set of experiments are performed with varying noise conditions. All the experimental studies are conducted for 3 different modeling paradigms, i.e., GMM-HMM, SGMM-HMM, and DNN-HMM. The linguistic evidence is captured using a bi-gram language model (LM) learned on the transcript of the acoustic training data. The same LM is employed for evaluating the recognition performances of all three systems.

In the first study, the acoustic models are trained and tested on clean speech. This is considered as the baseline system. Time-spliced MFCC feature vectors with LDA+MLLT+fMLLR transforms are used as input to the systems. The obtained results are shown in Table 5. It is observed that relative reduction of 3.81% and 5.53% in WER are observed in the case of SGMM-HMM and DNN-HMM-based systems with respect to the GMM-HMM-based system.

The next set of experiments are performed by appending the MFCC features with tone features and incorporating the tone model in all three kinds of acoustic modeling. The outcomes of this study are presented in Table 6. It is followed by testing of the developed systems on noisy test conditions. Two types of stationary noises, white and pink, are used for this purpose. SNR level of 5 dB is set for both types of noise. The results obtained in this study are also given in Table 6.

For clean test condition, a small but consistent improvement is noted with the addition of tonal information for all three

Table 5: *Baseline performances of three different kinds of acoustic modeling based CSR systems developed in this work*

Train Condition	Test Condition	Acoustic Model	%WER
Clean Speech	Clean Speech	GMM-HMM	13.37
		SGMM-HMM	12.86
		DNN-HMM	12.63

Table 6: *Assessment of the inclusion of tone features and tone modeling for Mizo continuous speech recognition under clean and noisy test conditions. The performances are evaluated separately for three different kinds of acoustic modeling techniques.*

Train Condition	Test Condition	Acoustic Model	%WER		
			MFCC	MFCC+ Tone Feats	MFCC+ Tone Feats+ Tone Model
Clean Speech	Clean Speech	GMM-HMM	13.37	13.35	13.12
		SGMM-HMM	12.86	12.80	12.29
		DNN-HMM	12.63	12.27	11.96
	White Noise (5dB)	GMM-HMM	60.31	42.48	42.47
		SGMM-HMM	59.08	40.83	34.66
		DNN-HMM	54.38	37.37	31.57
	Pink Noise (5dB)	GMM-HMM	65.42	55.49	53.27
		SGMM-HMM	65.67	51.11	46.33
		DNN-HMM	57.92	49.95	43.79

acoustic modeling cases. In case of 5 dB white noise test condition, it is observed that DNN-HMM-based modeling approach helps in reduction of WER from 60.31% in GMM-HMM-based system to 54.38% using MFCC features. The WER is reduced to 37.37% when tonal features are appended to the MFCC features. The WER is further reduced to 31.57% when tone modeling is incorporated in addition to the MFCC + tonal features. Similarly, for 5 dB pink noise test condition, relative reductions of 13.76% and 24.4% in WER are achieved with the inclusion of tone features and tone model respectively in the DNN-HMM system.

For the summarized assessment, the relative improvements obtained with/without the inclusion of tonal features as well as tone modeling for all the systems under different testing conditions are computed and are shown in Figure 2. It can be noted that under both noisy test conditions substantially larger relative improvements have resulted due to the inclusion of tonal information. This, in turn, supports our earlier argument that by exploiting the tonal characteristics present in the speech, a more noise robust CSR performance can be achieved.

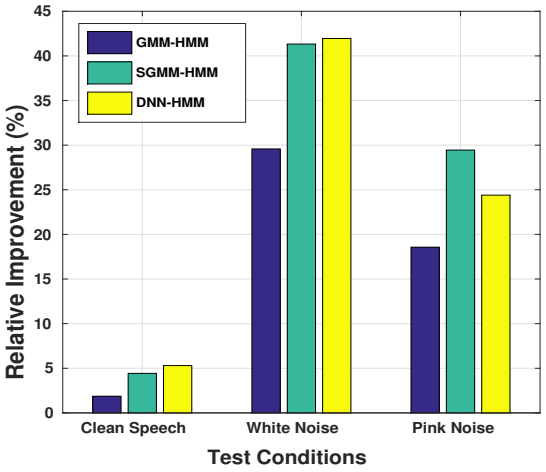

Figure 2: *Percentage relative improvement in WER obtained with inclusion of both tonal feature and tone modeling under different test conditions.*

6. Summary and conclusion

In this paper, we describe the attempts made for developing a Mizo CSR system. Since Mizo is a tonal language, the lexical meaning of a word changes with variation in the tone. For enhancing the recognition performance, the pitch related features and the tonal characteristics have been incorporated in Mizo CSR systems created following three different kinds of acoustic modeling approaches. The resulting tonal CSR systems have resulted in significantly improved recognition performances, in particular, under noisy test conditions. It is known that the pitch (fundamental frequency) information undergoes lesser degradation than the vocal tract information in presence of additive noise. This explains why the explored tonal Mizo CSR systems exhibit marked robustness for the test data corrupted with white and pink noise types. For the simplicity of considered noise types, a more detailed study involving realistic noise types is warranted and the same will be undertaken as the future work.

7. Acknowledgements

The speech corpus used in this work was developed for the project titled “Acoustic and Tonal Features based Analysis of Mizo”, funded by the Department of Electronics & Information Technology (DeitY), Ministry of Communication & Information Technology (MC&IT), Government of India.

8. References

[1] T. Lee, W. Lau, Y. W. Wong, and P. Ching, "Using tone information in cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, no. 1, pp. 83–102, 2002.

[2] Y. Qian, T. Lee, and F. K. Soong, "Use of tone information in continuous cantonese speech recognition," in *Speech Prosody 2004, International Conference*, 2004.

[3] W. Lau, T. Lee, Y. W. Wong, and P. Ching, "Incorporating tone information into cantonese large-vocabulary continuous speech recognition," in *Sixth International Conference on Spoken Language Processing*, 2000.

[4] Y. Cao, S. Zhang, T. Huang, and B. Xu, "Tone modeling for continuous mandarin speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2-3, pp. 115–128, 2004.

[5] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," in *Ninth International Conference on Spoken Language Processing*, 2006.

[6] N. Kertkeidkachorn, P. Punyabukkana, and A. Suchato, "Using tone information in thai spelling speech recognition," in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, 2014.

[7] H. Q. Nguyen, P. Nocera, E. Castelli *et al.*, "Using tone information for vietnamese continuous speech recognition," in *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*. IEEE, 2008, pp. 103–106.

[8] T. Drugman, Y. Stylianou, L. Chen, X. Chen, and M. J. F. Gales, "Robust Excitation-based Features for Automatic Speech Recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4664–4668.

[9] H. Yasui, K. Shinoda, S. Furui, and K. Iwano, "Noise robust speech recognition using spectral subtraction and f0 information extracted by hough transform," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, 2009, pp. 631–634.

[10] W. Lalhminghlui, P. Mazumdar, and P. Sarmah, "Behaviour of Tone Sandhi in Different Morphological Structures in Mizo," in *Proc. 23rd Himalayan Languages Symposium*, Tezpur University, Tezpur, 2017.

[11] W. Lalhminghlui and P. Sarmah, "Production and Perception of Rising Tone Sandhi in Mizo," in *Proc. TAL 2018*, Berlin, Germany, 2018.

[12] L. Chhangte, "Mizo Syntax," Ph.D. dissertation, University of Oregon, 1993.

[13] L. Fanai, "Some Aspects of The Lexical Phonology of Mizo and English: An Autosegmental Approach," Ph.D. dissertation, CIEFL, Hyderabad, India, 1992.

[14] P. Sarmah and C. R. Wiltshire, "A Preliminary Acoustic Study of Mizo Vowels and Tones," *J. Acoust. Soc. Ind*, vol. 37, no. 3, pp. 121–129, 2010.

[15] P. Sarmah, L. Dihingia, and W. Lalhminghlui, "Contextual Variation of Tones in Mizo." in *INTERSPEECH*, 2015, pp. 983–986.

[16] D. Govind, P. Sarmah, and S. M. Prasanna, "Role of Pitch Slope and Duration in Synthesized Mizo Tones," in *Proc. Speech Prosody 2012*, 2012.

[17] B. D. Sarma, P. Sarmah, W. Lalhminghlui, and S. M. Prasanna, "Detection of Mizo Tones," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] A. Weidert, *Componential Aanalysis of Lushai Phonology*. John Benjamins Publishing, 1975, vol. 2.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[20] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[21] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[22] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1043–1046.

[23] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian Mixture Models for Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 4330–4333.

[24] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2494–2498.

[25] R. Haeb-Umbach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. Washington, DC, USA: IEEE Computer Society, 1992, pp. 13–16.

[26] H. Abbasian, B. Nasersharif, A. Akbari, M. Rahmani, and M. Moin, "Optimized linear discriminant analysis for extracting robust speech features," in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*. IEEE, 2008, pp. 819–824.

[27] M. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[28] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, pp. 661–664.