INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Speech Transcription for Low-Resource Languages – The Case of Yoloxóchitl Mixtec (Mexico)

Vikramjit Mitra[1], Andreas Kathol[1], Jonathan D. Amith[2], Rey Castillo García[3]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2]Gettysburg College, PA, USA
[3]Secretaría de Educación Pública, State of Guerrero, Mexico
vikramjit.mitra@sri.com

Abstract

The rate at which endangered languages can be documented has been highly constrained by human factors. Although digital recording of natural speech in endangered languages may proceed at a fairly robust pace, transcription of this material is not only time consuming but severely limited by the lack of native-speaker personnel proficient in the orthography of their mother tongue. Our NSF-funded project in the Documenting Endangered Languages (DEL) program proposes to tackle this problem from two sides: first via a tool that helps native speakers become proficient in the orthographic conventions of their language, and second by using automatic speech recognition (ASR) output that assists in the transcription effort for newly recorded audio data. In the present study, we focus exclusively on progress in developing speech recognition for the language of interest, Yoloxóchitl Mixtec (YM), an Oto-Manguean language spoken by fewer than 5000 speakers on the Pacific coast of Guerrero, Mexico. In particular, we present results from an initial set of experiments and discuss future directions through which better and more robust acoustic models for endangered languages with limited resources can be created.

Index Terms: automatic speech recognition, endangered languages, large vocabulary continuous speech recognition, articulatory features, tonal features, acoustic-phonetic features, convolutional neural networks.

1. Introduction

1.1. Project goals

Documenting endangered languages offers great potential to contribute exceptional primary data for linguistic research. The rate at which endangered languages can be documented, however, has been highly constrained by human factors. This is particularly true in regard to corpus development: although digital recording of natural speech in endangered languages may proceed at a fairly robust pace, transcription of this material is not only time consuming but severely limited by the lack of native-speaker personnel proficient in the orthography of their mother tongue. Training of native-speaker transcribers by language experts requires considerable time, and typically few individuals are ever available for transcription work. This inevitably results in a transcription bottleneck that significantly slows documentation efforts and corpus-based lexico-semantic and morphosyntactic research. The present project addresses both shortcomings: (1) it provides for the training of native speakers in transcription, and (2) it builds automatic speech recognition (ASR) software (discussed in this study) for an endangered language and relies on those trainees for annotation, so that word-recognition accuracy may be improved over time.

1.2. Overview of Yoloxóchitl Mixtec

The project leverages existing ASR speech technologies trained on time-coded transcriptions of recordings from 24 speakers of Yoloxóchitl Mixtec (YM), an endangered Oto-Manguean language of fewer than 5,000 speakers, which is spoken in a few communities on the Pacific coast of Guerrero, 120 km east of Acapulco, Mexico. YM morphology is relatively simple with some inflectional (completive aspect, iterative/repetitive) and derivational (inchoative, denominalization through tone, causativization through prefixation and detransitivization through tonal alternations) features. Intervocalic lenition is prevalent. Unlike in other Mixtec languages, however, compounding seems to be relatively common. Person marking is by enclitics that often motivate vowel harmonization, stem-final elision of tone, and palatalization and labialization of stem-final consonants (tones are 1 low to 4 high).

1. *bi^{1}xi^{32}* (white hair, n.) > *ku^{3}-bi^{1}xi^{32}* (become white-haired)
2. *ku^{1}sũ^{1}* (sleep, v. intrans.) > *ndu^{3}- ku^{1}sũ^{1}* (return to sleep)
3. *yu^{1}u^{4}* (stone, n.) > *yu^{4}u^{4}* (solid, adj.) > *ku^{3}-yu^{4}u^{4}* (become)
4. *chi^{3}i^{3}* (become wet/moist) > *sa^{4}-chi^{3}i^{3}* (moisten)
5. *ka^{3}ã^{2}* (perforate, v. trans.) <> *ka^{1}ã^{1}* (perforate, v. intrans.)
6. *si^{13}su^{2}* (*Bauhinia* sp.) < *sĩʔ^{1}ĩ^{3}* ('foot') + *i^{3}su^{2}* ('deer')
   Note loss of nasalization and laryngealization.
7. *ndi^{1}ku^{4}chi^{4}* (broom) | [ndi^{1}ku^{4}či^{4}] [ndi^{1}gu^{4}či^{4}] [nd^{j}u^{14}či^{4}]
8. *be^{3}ʔã^{4}* | *be^{3}ʔe^{3}=ã^{4}* | house=3sgFem | her house
9. *ku^{1}s^{w}ĩ^{1}* | *ku^{1}sũ^{1}=i^{1}* | sleep (irreal)=1sg | I will sleep
10. *s^{y}uʔ^{3}un^{4}* | *siʔ^{3}i^{4}=ũ^{4}* | mother=2sg | your mother

Amith and Castillo García began work on YM in 2007 and have continued their efforts to the present, resulting in various publications, presentations, and manuscripts ([1], [2], [3], [4], [5], [6], [7]). In addition to some 70 hours of elicitation sessions targeting the phonetic and phonological study of nasalization and tone, their efforts have produced material that forms the basis for the present ASR effort: a 125-hour corpus of time-coded transcription of natural speech and a 2300-entry lexicon that comprises words both in the corpus and not yet recorded in natural speech. Though the material is substantive for endangered language documentation efforts, it is well within the low-resource range of language material usually used for natural language processing, including ASR. The project's objective is testing the potential of ASR technologies

http://dx.doi.org/10.21437/Interspeech.2016-546

to efficiently increase the size of transcribed corpora in low-resource endangered languages and, in the process, to suggest guidelines for endangered language documentation efforts that might seek to use ASR at some point.

1.3. Challenges for YM ASR

The extant YM time-coded transcriptions will be used as a gold standard to train native speakers to write YM. These speakers will then be able to annotate initial ASR outputs, correcting transcription hypotheses as needed and thus providing a mechanism to continually improve both recognition accuracy and native-speaker transcription skills. However, beyond the difficulties presented by a low-resource language, the phonology of YM is also particularly challenging for ASR. Not only does it manifest oral, nasal, and laryngealized vowels, but it has a complex lexical, inflectional, and derivational tonal system with nine basic tones (level, rising, and falling) and up to 21 contrasts on a single disyllabic, dimoraic word, a level of minimal contrasts that negatively affects accuracy. In contrast to other Mixtec languages, however, YM lacks one characteristic—tone sandhi—that would further complicate accurate word-level tone recognition. In sum, the present project has implications for endangered language research, ASR for low-resource languages, and the building of ASR systems for complex tonal languages.

2. Features

2.1. Features types

We investigated an array of features for training the YM acoustic model. The features included mel-filterbank (MFB) energies; gammatone filterbanks (GFBs); articulatory features (a.k.a. TVs) [8]; pitch and voicing features; and acoustic phonetic (APs) features. The MFB features were extracted by using the Kaldi speech recognition toolkit [9], with 40 MFBs extracted. The GFBs were extracted by using SRI International's implementation of a time-domain gammatone filterbank, which contained 40 channels that were equally spaced on the equivalent rectangular bandwidth (ERB) scale. For the acoustic features, the analysis window was 25.6 ms, with a frame rate of 10 ms. The GFBs used a 15th power root nonlinear compression.

Articulatory Features (AFs), articulatory motions from spontaneous speech, have been demonstrated by previous studies [10, 11] to provide robustness to speech recognition systems. In this previous work, we used a deep neural network (DNN) with four hidden layers containing 2048 neurons [12], to generate the articulatory features from the speech signal. The speech signal was parameterized as amplitude modulation features (NMCCs, see [13] for more detail regarding NMCC computation) that were fed as input to the DNN acoustic-to-articulatory speech-inversion model (details regarding the model are presented in [12]). The DNN generated time-domain vocal tract constriction variables (TVs). The TVs provide kinematic information regarding vocal tract constriction location and degree during speech production [14]. For each input acoustic feature frame, the DNN generated an eight-dimensional vector, whose elements represented tongue tip constriction degree and location; tongue body constriction degree and location; labial aperture and protrusion; glottal opening-closing; and velic opening-closing.

Acoustic Phonetic (AP) features [15] represent acoustic-phonetic information (e.g., formant information, mean Hilbert envelope, and periodic and aperiodic energy in subbands [16]) and were analyzed at a 5 ms frame rate with a 10 ms analysis window. Thirteen APs were selected to represent information such as reflection coefficients, mean Hilbert envelope, periodic energy, aperiodic energy [16], and nasal energy [17]. Data such as periodic energy, Hilbert envelope, and other features provide information regarding voice quality and energy contour, among others.

The **Kaldi Pitch** tracker [18] comes with the Kaldi pitch recognition toolkit [9] and provided two-dimensional output consisting of pitch tracks and a normalized cross-correlation function that gave an indication about voicing information.

The **SAcC** pitch feature (for Subband Autocorrelation Classification) [19] is a noise-robust pitch tracker. SAcC involves a multilayered perceptron (MLP) classifier trained on subband autocorrelation features to estimate, for each time frame, the posterior probability over a range of quantized pitch values and one "no-pitch" output.

The third pitch tracker used in our experiments is known as the **MBCombF0** pitch-tracker, which is a modification of the correlogram-based F0 estimation algorithm described in [20]. In MBCombF0, a frame length of 100 ms was used, where the speech signal is downsampled to 8 kHz and split into four subbands that cover 0 to 3.4 kHz. Each subband had a 1 kHz bandwidth and overlapped the adjacent filter by 0.2 kHz. Envelope extraction was then performed on each subband stream, followed by multichannel comb-filtering with comb filters of different inter-peak frequencies. Next, reliable comb-channels were selected individually for each subband by using a three-stage selection process (more details are presented in [20]). At the final step, MBCombF0 processing generated four subband summary correlograms that were combined by using a subband reliability weighting scheme to form the multiband summary correlogram. Time smoothing was applied to the multiband summary correlogram [21], and the resulting information was used to generate the MBCombF0 voicing and pitch feature used in this work.

3. Acoustic model

We trained different acoustic models for the YM speech recognition task, for which we explored traditional DNNs; convolutional neural nets (CNNs); time-frequency convolutional nets (TFCNNs) [22]; and the recently proposed hybrid convolutional neural net (HCNN). The acoustic models were trained with the features described in the previous section.

It was shown in [23] that CNNs give lower WERs compared to DNNs when using filterbank features for ASR tasks, and that GFBs performed as well as or better than the MFBs. Hence, in this study, we used the MFB and GFB DNN/CNN model as our baseline systems.

To generate the alignments necessary for training the CNN system, a GMM-HMM model was used to produce the labels for the senones. Altogether, the GMM-HMM system produced 1644 context-dependent (CD) states for the YM data. The input features to the acoustic models were formed by using a context window of 15 frames (7 frames on either side of the current frame).

The acoustic models were trained by using cross-entropy on the alignments from the GMM-HMM system. For the

CNN, 200 convolutional filters of size 8 were used in the convolutional layer, and the pooling size was set to 3 without overlap. The subsequent fully connected network had four hidden layers, with 1024 nodes per hidden layer, and the output layer included as many nodes as the number of CD states for the given dataset. The networks were trained by using an initial four iterations with a constant learning rate of 0.008, followed by learning rate halving based on cross-validation error decrease. Training stopped when no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed by using stochastic gradient descent with a mini-batch of 256 training examples. For the DNN systems, we used five layers with 1024 neurons in each layer, with similar learning criteria as the CNNs.

The TFCNN architecture is based on [22], for which two parallel convolutional layers are used at the input, one performing convolution across time, and the other across the frequency scale of the input filterbank features. That work showed that the TFCNNs gave better performance compared to their CNN counterparts. Here, we used 75 filters to perform time convolution and 200 filters to perform frequency convolution. For time and frequency convolution, eight bands were used. A max-pooling over three samples was used for frequency convolution, while a max-pooling over five samples was used for time convolution. The feature maps after both the convolution operations were concatenated and then fed to a fully connected neural net, which had 1024 nodes and four hidden layers.

The HCNN is a modified deep neural network architecture to jointly model the acoustic and the articulatory space [12]. In the HCNN, two parallel neural networks are trained simultaneously. These two parallel neural networks model two processes: (1) the learning of the acoustic space through a time-frequency convolutional net and (2) the learning of a temporal feature trajectory space through a time convolutional net. These two convolution layers had the same parameter specification as that used in the TFCNNs. The time-convolution layer contained 30 filters, followed by a max-pooling over five samples. The fully connected DNN layers were different in size, with 1024 neurons used for the TFCNN, and 512 neurons used for the time convolutional net. Note that both parallel networks were jointly trained.

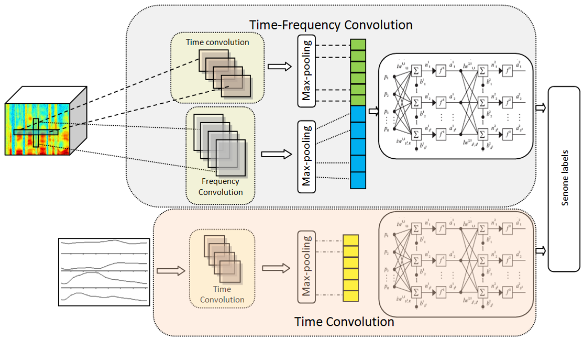

Figure 1: *Schematics of the hybrid convolutional neural network (HCNN). The top layer represents a TFCNN whose input is typically filterbank features, and the bottom layer represents the temporal features processed through time convolution.*

4. Results

Two baseline acoustic models were trained by using five-hidden-layered DNNs with MFB and GFB features. Each hidden layer contained 1024 neurons. A trigram language model (LM) was used to decode the ASR hypothesis. The word error rates (WERs) from the baseline systems are shown in Table 1.

Table 1. *Baseline WERs from MFB and GFB features for YM DNN acoustic models.*

Features	WER (%)
MFB	36.9
GFB	35.0

Table 1 shows that the GFB baseline system gave lower error rates than the MFB features. GFB features have demonstrated better robustness to background distortions [23], and given that the YM recordings were made with realistic background conditions, the recordings can be expected to contain ambient noise and recording-microphone distortions. MFBs are typically found to be sensitive to background acoustic distortions, and hence Table 1 shows that GFBs under similar background conditions performed better than the MFBs, thus resulting in lower WERs for YM speech recognition.

Next, we focused on using the GFB features, adding articulatory features, pitch features, and acoustic phonetic features. We then retrained the DNN acoustic model. Table 2 shows that using additional features with GFBs reduced the WER by 6.86% relative to the GFB baseline DNN system.

Table 2. *WERs from GFB features with articulatory features, pitch features, and acoustic phonetic features added.*

Features	WER (%)
GFB	35.0
+articulatory	34.9
+KF0	32.9
+SAcC	32.6
+MBCOMBF0	32.5
+AP	32.6

In addition to the DNN systems, we also explored using CNN and the recently proposed HCNN system using both GFBs and articulatory features. Table 3 shows the results from using the CNN and the HCNN systems.

Table 3. *WERs from GFB-CNN and GFB+articulatory HCNN systems.*

Features	Model	WER (%)
GFB	CNN	32.9
GFB	TFCNN	32.3
GFB+articulatory	HCNN	32.4

Table 3 shows that using the TFCNNs and the HCNNs both reduced the WERs compared to the GFB-CNN model, indicating that leveraging temporal information and articulatory information was useful in performing more accurate speech recognition for the YM data.

To compare how much tonal contrasts impact the performance of YM speech recognition, we removed all tonal contrasts from the YM references and retrained the DNN acoustic model and the LM. Table 4 shows the results from the GFB TFCNN systems after tonal contrast removal, with a 4.3% relative improvement in WER after removal of the tonal contrasts from the references.

Table 4. *WERs from GFB-TFCNN system with and without tonal contrasts.*

Features	Tonal contrast	WER (%)
GFB	Yes	32.3
GFB	No	30.9

Table 2 shows that the AP features may not be providing useful information for YM ASR, hence using these features resulted in an increase in WER. Tables 2 and 3 also illustrate that by using convolutional acoustic models, such as TFCNNs and HCNNs, we can obtain performance as good as that obtained from using acoustic+articulatory+pitch features. This finding indicates that further improvement in performance could be achieved if the acoustic+articulatory+pitch feature combination were used in a CNN or other advanced acoustic models.

We investigated using the HCNN for acoustic+articulatory+pitch features, where the acoustic features were input to a time-frequency convolutional net, and the articulatory+pitch feature was fed to a time convolutional net that shared their output layer. In addition, we explored bottleneck features, in which a five-hidden-layer DNN acoustic model was trained in a supervised manner with a bottleneck layer of 40 neurons. Figure 2 shows the architecture for bottleneck-feature extraction and training a DNN acoustic model.

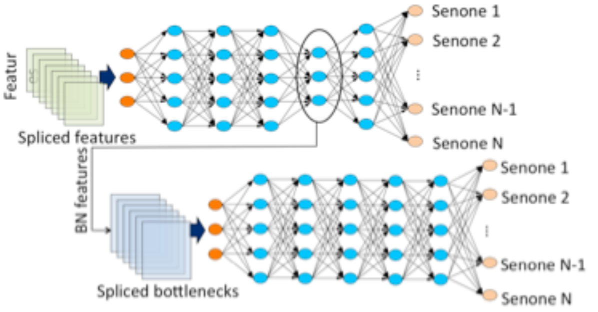

Figure 2: *Bottleneck-feature extraction and DNN acoustic model training. Note that the senones in the figure represent context-dependent triphone states.*

We trained a bottleneck (BN) layer with 40 neurons, and the bottleneck features were spliced (contextualized) with 15 frames. DNN acoustic models (we name these models "BN-DNN") with six hidden layers were trained having 2048 neurons. We also trained DNN acoustic models with acoustic+articulatory+pitch features to compare their performance with respect to the bottleneck features. In addition, we also explored using the HCNN network, in which GFB was used to train the time-frequency convolution part of the HCNN, and articulatory+pitch features to train the time convolution part of the HCNN. The results are shown in Table 5.

Table 5. *WERs from BN-DNN, multi-feature-fused DNN, and HCNN systems with different network sizes.*

Features	Model	WER (%)
GFB+Artic.+Pitch	DNN	31.1
GFB+Artic.+Pitch	HCNN	31.4
BN	DNN	31.3

Table 5 shows that the multi-feature DNN gave the best performance, with a WER of 31.1%, which is as-good-as the performance of the GFB-TFCNN system without tonal contrasts reported in table 4. Note that the DNN models in Table 5 had six hidden layers with 2048 neurons, whereas the HCNN had 5 hidden layers with the same number of neurons. This finding indicates that fusing multiple features such as acoustic, articulatory and pitch features was beneficial for creating a better acoustic model for YM.

5. Conclusion

Although training an acoustic model for YM data was difficult given the diversity of tonal contrasts that exist in the language, this study demonstrates that a reasonable acoustic model can be trained. The best WER obtained from the acoustic models presented in this study was close to 30% and, when all the tonal contrasts were removed, was almost as good as that of the baseline system trained for the language. We observed that the GFB features performed better than the MFB features, which may be due to the nonlinear root compression of the GFBs, which are known to be more robust to acoustic and noise distortions. Note that the YM data was gathered from a field collection and, hence, contains varying background distortions and recording conditions. We found that robust acoustic models such as HCNN and CNNs gave reasonable performance improvement over the baseline system trained with DNN systems. The best performance, however, was obtained from using a simple DNN acoustic model trained with multiple acoustic features.

In the future, we plan to investigate unsupervised learning of BNs through an autoencoder network that has already shown some interesting gains for ASR and comparable performance with respect to DNN-BN systems [24]. We will also explore late fusion of multiple systems [25], which typically has shown impressive performance improvements in the literature.

6. Acknowledgement

Direct support for this project came from the Documenting Endangered Languages program, NSF awards 1500738 and 1500595: "Collaborative Research: Speech Technology-Enhanced Annotation and Training Tool for Yoloxóchitl Mixtec (xty)," Andreas Kathol and J. Amith, co-PIs. We also acknowledge and thank the following institutions for their support for Yoloxóchitl Mixtec documentation: Ford Foundation grant 1075-0256; School of Oriental and African Studies awards PPG0048 and MDP0201; National Science Foundation awards 0966462 and 1435831.

7. References

[1] E. Palancar, J. D. Amith, and R. Castillo García, “Verbal inflection in Yoxolochitl Mixtec,” in E. Palancar and J.L. Léonard, eds., *Tone and Inflection: New Data under New Perspectives*, Berlin, Germany: De Gruyter Mouton, pp. 295–336, 2016.

[2] C. DiCanio, H. Nam, J. D. Amith, R. Castillo García, D. H. Whalen, “Vowel variability in elicited versus running speech: Evidence from Mixtec,” *Journal of Phonetics: Special Issue on the Impact of Stylistic Diversity on Phonetic and Phonological Evidence and Modeling* vol. 48, pp. 45–59, 2015.

[3] C. DiCanio, J. D. Amith, and R. Castillo García, “The phonetics of moraic alignment in Yoloxóchitl Mixtec,” *Proceedings of the Fourth International Symposium on Tonal Aspects of Languages*, Nijmegen, The Netherlands, May 13–16 May, 203–210, 2014.

[4] C. DiCanio, H. Nam, D. H. Whalen, H. T. Bunnell, J. D. Amith, and R. Castillo García, “Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment,” *Journal of the Acoustical Society of America* vol. 134, no. 3, pp. 2235–2246, 2013

[5] C. DiCanio, H. Nam, D. H. Whalen, H. T. Bunnell, J. D. Amith, and R. Castillo García, “Assessing agreement level between forced alignment models with data from endangered language documentation corpora,” in INT*ERSPEECH 2012 - 13th Annual Conference of the International Speech Communication Association*, September 9–13, Portland, Oregon, USA, Proceedings.

[6] R. Shosted, J. D. Amith, R. Castillo García, and C. DiCanio, “Nasalization and voiceless obstruents in Yoloxóchitl Mixtec: An aerodynamic analysis,” Paper presented at the *Workshop on the Sound Systems of Mexico and Central America,* April 4–6, Yale University, New Haven, Connecticut, 2014.

[7] J. D. Amith and R. Castillo García, “Dictionary of Yoloxóchitl Mixtec,” unpublished ms.

[8] Goldstein, “Retrieving tract variables from acoustics: A comparison of different machine learning strategies,” *IEEE Journal of Selected Topics on Signal Processing, Sp. Iss. on Statistical Learning Methods for Speech and Language Processing*, vol. 4, iss. 6, pp. 1027–1045, 2010.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.

[10] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, “Articulatory information for noise robust speech recognition,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, iss. 7, pp. 1913–1924, 2010.

[11] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson and E. Saltzman, “Articulatory features from deep neural networks and their role in speech recognition,” in *Proc. of ICASSP*, pp. 3041–3045, Florence, 2014.

[12] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman and M. Tiede, “Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition,” under review, 2016.

[13] V. Mitra, H. Franco, M. Graciarena and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition,” *Proc. of ICASSP*, pp. 4117–4120, Japan, 2012.

[14] H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman and L. Goldstein, “A procedure for estimating gestural scores from natural speech,” *Journal of the Acoust. Society of America*, vol. 132, iss. 6, pp. 3980–3989, 2012.

[15] A. Juneja, “Speech recognition based on phonetic features and acoustic landmarks,” PhD thesis, University of Maryland College Park, December 2004.

[16] O. Deshmukh, J. Singh, C. Espy-Wilson. 2004. “A novel method for computation of periodicity, aperiodicity and pitch of speech signals,” *Proceedings of the 34th International Conference on Acoustics, Speech and Signal Processing*, 17–21 May, Montreal, Canada, pp. 117–20.

[17] [10] T. Pruthi and C. Espy-Wilson, “Acoustic parameters for the automatic detection of vowel nasalization,” *Proceedings of INTERSPEECH*, pp. 1925–1928, 2007.

[18] [11] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpu, “A pitch extraction algorithm tuned for automatic speech recognition,” in *Proc. of ICASSP*, 2014.

[19] B. S. Lee and D. Ellis, “Noise robust pitch tracking by subband autocorrelation classification,” in *Proc. of Interspeech*, 2012.

[20] L. N. Tan and A. Alwan, “Noise-robust F0 estimation using SNR-weighted summary correlograms from multi-band comb filters,” in *Proc. ICASSP*, pp. 4464–4467, 2011.

[21] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Proc. of ICASSP*, pp. 1331–1334. 1997.

[22] V. Mitra and H. Franco, “Time-frequency convolution networks for robust speech recognition,” in *Proc. of ASRU* 2015.

[23] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels and M. Graciarena, “Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions,” in *Proc. of Interspeech*, pp. 895–899, Singapore, 2014.

[24] V. Mitra, C. Bartels, M. Graciarena, J. van Hout, H. Franco and D. Vergyri, “Unsupervised adaptation of deep neural networks to unseen and noisy channel conditions,” submitted for review.

[25] J. G. Fiscus, “A post-processing system to yield reduced word error rates: recognizer output voting error reduction. (ROVER),” *Proc. of ASRU*, pp. 347–354, 1997.