HIGHLY ACCURATE PHONETIC SEGMENTATION USING BOUNDARY CORRECTION MODELS AND SYSTEM FUSION

Andreas Stolcke[1] *Neville Ryant*[2] *Vikramjit Mitra*[3] *Jiahong Yuan*[2] *Wen Wang*[3] *Mark Liberman*[2]

[1]Microsoft Research, Mountain View, CA, USA
[2]University of Pennsylvania, Philadelphia, PA, USA
[3]SRI International, Menlo Park, CA, USA

ABSTRACT

Accurate phone-level segmentation of speech remains an important task for many subfields of speech research. We investigate techniques for boosting the accuracy of automatic phonetic segmentation based on HMM acoustic-phonetic models. In prior work [25] we were able to improve on state-of-the-art alignment accuracy by employing special phone boundary HMM models, trained on phonetically segmented training data, in conjunction with a simple boundary-time correction model. Here we present further improved results by using more powerful statistical models for boundary correction that are conditioned on phonetic context and duration features. Furthermore, we find that combining multiple acoustic front-ends gives additional gains in accuracy, and that conditioning the combiner on phonetic context and side information helps. Overall, we reduce segmentation errors on the TIMIT corpus by almost one half, from 93.9% to 96.8% boundary accuracy with a 20-ms tolerance.

Index Terms— *phonetic segmentation, phone boundary model, forced alignment, HMM, regression, system fusion.*

1. INTRODUCTION

The availability of large corpora of transcribed speech data has been an important enabler for achieving the current level of performance in automatic speech recognition (ASR). However, ASR training only requires transcription at the word-level, whereas research fields such as phonetics, sociolinguistics, and psychology, depend on accurate phone-level segmentations and transcriptions. Manual phonetic segmentation is time-consuming and expensive, taking up to 400 times real time [1] or 30 seconds per phone [2]. Automatic phonetic segmentation is much needed.

A common approach for automatic phonetic segmentation is "forced alignment", based on two inputs: a speech audio waveform and a transcription at the phone- or word-level. In the case of word-level transcription, the words are first mapped into a phone sequence using a combination of pronouncing dictionary and grapheme-to-phoneme rules. Acoustic phone models in the hidden Markov model (HMM) framework are then trained, as would be required for a phone recognizer [3][4][5][6][7][8]. In this approach, each phone is a HMM of typically 3-5 states, and the entire utterance is modeled by the concatenation of phone HMMs (in the case of alternate pronunciations, a lattice of states may be used). Each state has a self-loop to emit a variable number of speech frame feature vectors, computed by an acoustic front end. The sequence of frames comprising a speech utterance is aligned to the known phone sequence or lattice by finding the sequence of hidden states that maximizes the utterance likelihood under the state acoustic models. On the standard TIMIT corpus, the reported performances of conventional HMM-based forced alignment systems range from 80%-89% agreement (of all boundaries) within 20 ms compared to manual segmentation [6].

A main drawback of the HMM-based forced alignment for phonetic segmentation is that phone boundaries are not represented in the model and the training procedure does not explicitly optimize for boundary accuracy. The boundaries are merely a by-product of the likelihood maximization over possible state alignments. Contrast this with the manual phonetic segmentation process, in which the acoustic landmarks at the phone boundaries [9], e.g., an abrupt spectral change, are used to determine the location of a boundary. One method to address this shortcoming is discriminative training of alignment models, as in [14]. Another approach is the use of features specifically designed to model boundary events, such as energy-based features and distinctive phonetic features, and the use of observation-dependent state transition probabilities [15]. That system achieved 93.36% agreement within 20 ms on TIMIT compared to manual segmentation, a result that comes close to the average agreement of 93.49% among human labelers [6][15].

Another group of approaches uses a two-stage architecture, where HMM-based forced alignment generates rough boundary estimates that are then refined by more specialized (and discriminatively trained) models. For example, [10] used energy changes in different frequency bands for boundary correction, [11] trained support vector machine (SVM) classifiers to differentiate boundaries from non-boundary positions, and [12] and [13] employed neural network to refine phone boundaries.

In our own prior work [26], we achieved state-of-the-art alignment accuracy on TIMIT (93.92%) by combining explicit boundary models (using HMM states associated with transitions between phones), carefully choosing context-dependency (monophone models can outperform triphone models), constraining phone model training by human boundary labels, and employing simple post-alignment correction models (constant or linear adjustment based on training data statistics for each phone label pair).

In this paper we build on the above work to achieve further improvements of the state-of-the-art in TIMIT phone alignment accuracy. First, we consider more sophisticated boundary

correction models than used previously. We add a wider array of local features (phonetic context, phone durations, etc.) and apply general regression model architectures, including regression trees and neural nets. We also investigate the use of global speaker-dependent features. Second, we instantiate our architecture with a large array of acoustic front ends, comparing performance to standard MFCC and PLP versions of our system. Finally, we investigate combinations of multiple systems based on diverse front ends, an approach that has been very successful in ASR systems [27]. A static combination of multiple alignment outputs has been proposed by [28]; here we develop an approach that allows the combiner model to learn from contextual features which subsystems to trust.

In the following we first introduce the data set and the evaluation method (Section 2). In Section 3 we summarize out the basic boundary estimator developed in [26], as well as the acoustic front ends employed. Section 4 investigates the various boundary correction models. Section 5 compares performance for different front ends and ways to combine them, followed by conclusions.

2. DATA AND EVALUATION

The TIMIT corpus was used [22]. Excluding the "dialect calibration" sentences (SA sentences), 3,696 utterances from the training partition of the corpus were used for training and 1,344 utterances from the test partition were used for testing. As listed in Table 1, the 61 TIMIT phonemes were mapped to 54 phonemes (following [15], pp. 357). The boundaries between two pauses, including stop closures, were excluded from evaluation. There were 136,450 boundaries in the training set, and 49,248 boundaries in the test set for evaluation.

The accuracy of automatic segmentation is measured in terms of what percentage of the automatically labeled boundaries are within 20 ms of the manually labeled boundaries, which has been widely used in previous studies.

Table 1: The phoneme set (54 phonemes)

Pauses and stop closures	/pau/, /pcl/, /bcl/, /tcl/, /dcl/, /kcl/, /gcl/
Vowels	/aa/, /ae/, /ah/, /ao/, /aw/, /ax/, /axh/, /axr/, /ay/, /eh/, /er/, /ey/, /ih/, /ix/, /iy/, /ow/, /oy/, /uh/, /uw/, /ux/
Glides	/l/, /r/, /w/, /y/, /hh/, /hv/
Nasals	/m/, /n/, /ng/, /nx/
Plosives	/b/, /d/, /g/, /p/, /t/, /k/, /dx/, /jh/, /ch/
Fricatives	/s/, /z/, /sh/, /zh/, /f/, /v/, /th/, /dh/

3. SYSTEM COMPONENTS

3.1 Phone and boundary modeling

Both monophone and phone boundary HMMs were trained, using the HTK toolkit [25]. Stops, stop closures, the vowel /axh/ ("devoiced schwa"), nasals, and liquids (/l/, /r/) are 1-state HMMs; the "true" diphthongs (/ay/, /aw/, /oy/) are 5-state HMMs; and the other phonemes are 3-state HMMs. The phone boundary HMMs are a special 1-state HMM, in which the transition probabilities $a_{01} = 1$, $a_{11} = 0$, and $a_{12} = 1$ (as shown in Figure 1). Therefore, a boundary can have one and only one state occurrence, i.e., aligned with only one frame.

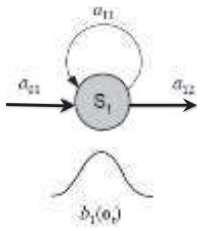

Figure 1: The boundary model is a special 1-state HMM for phone boundaries with transition probabilities $a_{11} = 0$ and $a_{12} = 1$.

The special 1-state phone boundary HMMs were trained using frames extracted at the manually labeled phone boundaries, one frame for each boundary. Within-word and cross-word boundaries were differentiated. The full set of boundary models contained 5,832 states, one state for each boundary type (54 phonemes on the left × 54 phonemes on the right × 2). The boundary states were tied through decision-tree based clustering, similar to triphone state tying. The monophone HMMs were trained using frames extracted within the manually labeled phone boundaries, excluding the boundary frames. Each state has 8 Gaussian mixture components with diagonal covariance matrices.

In testing, forced alignment was run over utterances representing their phone transcriptions (including the boundary labels derived from the phone identities). The forced alignment boundaries were then adjusted by applying statistical models that are conditioned on phonetic context and duration features, built on the training data.

3.2 Acoustic front ends

The baseline system [26] was built on Perceptual Linear Prediction (PLP) features implemented by HTK. One of the goals of this study was to explore and combine a variety of acoustic front ends, as is common in ASR systems. We included systems based on the traditional Mel-frequency cepstral coefficients (MFCCs), and PLP features using RASTA processing (RASTA-PLP) [29].

The first group of nonstandard features explored in our study are perceptually motivated features that have been shown to have significant robustness to noise and channel degradation. Normalized Modulation Cepstral Coefficient (NMCC) [30] is a noise robust acoustic feature obtained from tracking the amplitude modulations (AM) of gammatone-filtered subband speech signals in time domain. The AM estimates are obtained using the discrete energy separation algorithm (DESA) [31] based on the nonlinear Teager's energy operator [32]. The modulation information after root power compression is used to create a cepstral-like feature, where the first thirteen discrete cosine transform (DCT) coefficients were retained and their dynamic coefficients were also computed. Modulation of Medium Duration Speech Amplitudes version 1 (MMeDuSA-1) is a variant of NMCC features that uses medium-duration analysis windows (~51 ms) to estimate approximate AM information of band-limited speech signals [33]. MMeDuSA-2 [34] is similar to MMeDuSA-1[33] except that it has three extra dimensions to capture the summary speech modulation across frequency channels; these extra dimensions provide prosodic information and information about speech activity. Power Normalized Cepstral

Coefficients (PNCC) [35] is a feature obtained from gammatone filterbank analysis of speech signals, where the filterbank energies are power-normalized, bias-subtracted, root-compressed and then cosine-transformed to yield cepstra-like features. Mean Hilbert Envelop Cepstra (MHEC) [36] is a feature that uses gammatone filterbank to analyze speech and performs DCT on the mean Hilbert envelopes to construct a cepstra-like feature. Synchronized damped oscillator cepstral coefficients (SyDOCC) [37] are a set of perceptually motivated features that represent auditory hair cells as a set of damped oscillators excited by a set of time-synchronized band-limited speech signals.

We also explored a set of articulatory features originally developed for ASR [37]. The use of articulatory features is motivated by the fact that they are robust to coarticulation effects and have also shown some degree of robustness to channel and noise degradation [38]. The articulatory features (a.k.a. tract variables or TVs) define vocal tract constrictions dynamically in time. We fused the modulation information of the TVs with MFCCs and, after dimensionality reduction based on principal component analysis (PCA), created a 30-dimensional composite feature, which we call MFCC_ModTV-PCA [38].

Finally, we also explored a feature obtained from training an autoencoder (AE) network with 150 neurons using the NMCCs as input. An AE network maps the input space to itself and, as a consequence, its hidden variables learn the broad phonetic inventory of the acoustic space in an unsupervised manner. Once trained, the output (or generation) layer is split off from the network and the 150-dimensional hidden variables from the first (or extraction) layer are used as the AE feature. We applied PCA on these features to reduce them to 53 dimensions, ensuring that at least 95% of information is retained. Note that all the cepstra-like features used the first 13 DCT coefficients and their Δ, Δ^2, and Δ^3 information, resulting in feature vectors of dimension 52. In the case of MMeDuSA-2, the additional 3 dimensions resulted in a feature vector of dimension 55.

4. BOUNDARY CORRECTION MODELS

As discussed earlier, HMM segmentation points obtained from forced alignment (FA) need to be adjusted to achieve good boundary accuracy. In prior work [26], we used one of two statistical correction procedures, one for the boundaries between vowels or glides, and one for all other boundaries. The boundaries between vowel and glide phonemes are inherently subjective and labeling guidelines for such boundaries employ a heuristic rule by which "one-third of the vocalic region is assigned to the semivowel." We therefore built a linear model for vowel/glide phonemes that predicts the manually labeled boundary positions from the forced alignment positions of the two phonemes (phoneme center positions), the identities of the boundaries (the phonemes preceding and following the boundary), and the forced alignment boundary positions. The model was trained on the training portion of the data and applied to the test data. For all other boundaries, the mean difference between manually labeled and forced alignment boundaries for every boundary identity was calculated using the training data, and the forced alignment boundaries in the test set were shifted by these boundary-dependent time differences.

For this paper, we employed three more general regression models to predict true phone boundaries from FA boundaries. The first model consists of standard regression trees, as implemented by the M5PrimeLab package [39]. Input features to the model are the identities of the left/right phones and their durations (according to FA). The predicted variable is the signed deviation between FA and reference boundary. The phone identities are encoded either as unary feature vectors, or as a vector of 28 binary articulatory features. (We also tried M5' tree models [39], but found them to work slightly less well than standard regression trees.) The articulatory encoding works best for this model.

The second regression model was a simple multi-layer perceptron (neural network) with a single hidden layer of 20 units; the nonlinear hidden-unit activation function used was tanh. The training criterion was mean-squared error, and 30% of the training set was used for cross-validation. The input features were the same as described earlier for the regression tree model. We also investigated adding speaker-dependent features such as age, sex, and dialect region, but found no gains in preliminary experiments, and did not include these in results reported below.

The final model for boundary correction was a neural net using expanded input features and modified network topology, trained with more sophisticated optimization techniques. A network with one hidden layer of 128 rectified linear units [40] was trained for 2000 epochs (epoch = 250000 instances) using stochastic gradient descent with minibatches of size 128 and dropout of 50% in the hidden layer [41]. A decaying learning rate was used with initial learning rate set to 1. Momentum was increased from 0.25 to 0.5 linearly over the course of learning. Input features consisted of left and right predicted phone durations (as before), left and right phone identity coded using a 1-of-k scheme, and distance from the system boundary to beginning and end of the utterance (the distance features had not yielded improved predictions with the simpler regression models).

Table 2: Boundary accuracy (20ms tolerance) with different regression models

Correction model	Accuracy (%)
Baseline [26]: without corrections	91.02
Baseline [26]: with corrections	93.92
Regression tree	93.90
Neural network 1	94.17
Neural network 2	94.39

Table 2 shows the boundary accuracy within 20 ms for the various correction models. Also shown is the accuracy without correction of the underlying forced alignments, which were obtained with the system based on a single PLP front-end as described in [26]. Boundary correction by regression tree achieves the same accuracy as the simple heuristic correction model from that earlier work. The neural network models do improve on the previous result, reducing error by 4.1% and 7.7% relative, respectively. It is maybe surprising that the simple baseline model with linear heuristic correction is not far behind the completely data-driven models that use a wider array of input features. But the comparison between the two neural network models shows that the regression models may be data-starved and require very careful optimization.

5. SYSTEM COMBINATION

In addition to the UPenn PLP system used previously [26], we trained separate HMM systems for all the front ends described in Section 3.2, based on feature extractors implemented at SRI. Note that the PLP (HTK) features were zero-mean normalized, whereas

the SRI features, including PLP, did not employ utterance-level normalization.

Table 3: Boundary accuracy by front end, using neural network boundary correction model 1

Front end	Accuracy (%)
PLP (HTK)	94.17
PLP (SRI)	93.83
SyDOCC	93.73
MMeDuSA-2	93.71
PNCC	93.69
MMeDuSA-1	93.66
NMCC	93.57
MFCC	93.09
MFCC_ModTV-PCA	93.09
MHEC	92.75
AE	91.90

As shown in Table 3, all front ends, with the exception of AE, resulted in accuracies within 1.5% of the original system. The AE system was over 2% worse, and was not included in combination experiments. In comparing results of the different systems it is important to note that the model parameters (e.g., the number of mixture components, the number of hidden states, and the degree of boundary state tying) were tuned using HTK PLPs only, and then applied to all features. This expedient seemed justified given that our focus was system fusion, not the comparison between particular front ends.

As a baseline combiner approach, we can simply average boundary estimates from all systems. However, we want the combiner to learn which systems to trust as a function of context. To achieve this we trained a neural network with two hidden layers of 128 rectified linear units that takes the individual boundary estimates (we encode these as differences from the first system's prediction), along with articulatory features encoding one phone of context on each side of the boundary as input. The network is trained to estimate the deviation from the first input system's prediction to the true boundary time (i.e., a correction to a correction). The network was trained for 200 epochs (250000 instances per epoch) using stochastic gradient descent with minibatches of size 128 and 10% dropout in the hidden layers. Learning rate and momentum were kept constant at 1 and 0.5 respectively.

Table 4: Boundary accuracy for different system combination approaches

Combiner model	Accuracy (%)
Average	95.10
Neural net w/o phone context	96.39
Neural net with phone context	96.66
+ average of regression-tree systems	96.77

We found that the outputs of the simpler neural network 1 correction models (see Section 4) gave slightly better result than the more complex neural network 2 after combination, possibly because they were less correlated. (The optimization of individual system with respect to the combined result is a problem for future work.)

Results in Table 4 show that the neural net achieves a 26.3% relative error reduction over the average of systems. Providing phone context to the combiner network reduces error by an additional 7.5% relative.

A final improvement can be obtained by combining not just different front ends, but also different boundary correction models. To avoid doubling the number of input parameters to the combiner, we take an average of versions of all the systems using a regression tree backend instead of the neural net. This yields an accuracy of 94.99% (slightly worse than 95.10%, the average of systems with neural net backends). This average was then added as one additional input to the overall combiner, which was itself neural network-based. This final combination (last row of Table 4) gives an additional 3.3% relative error reduction. This overall combined system has 42.4% lower error than the best single boundary estimation system, and 46.9% lower than the best previous result (cf. Table 2).

We might also be interested in how error is reduced as a function of the number of available input systems. Figure 2 plots the drop in error as we added more systems to the simple averaging combiner, in the order of their individual accuracy (as shown in Table 3). As shown, there is some leveling off after the three best systems are combined, but incremental improvements are achieved even with ten front ends.

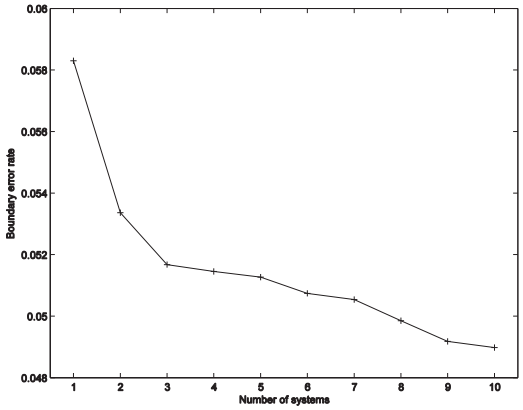

Figure 2: Boundary error rate as a function of number of front-ends/systems (combining by averaging)

6. CONCLUSIONS

We have investigated ways to improve phone boundary time estimates based on HMM forced alignments. Two sources of improvement are better correction models, namely neural networks that take phonetic context and duration features as input to estimate the deviation of the true phone boundary from the one given by Viterbi alignment. A further improvement in accuracy is achieved by combining estimates based on multiple acoustic feature front ends. A simple averaging is effective as long as all systems are roughly equal in quality, but here, too, neural networks that take phonetic context into account can yield additional gains. Overall, we achieve a substantial improvement in phone segmentation accuracy on TIMIT data, cutting the residual error rate (at 20 ms tolerance) almost in half compared to the previous best result.

6. ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-0964556. The articulatory features were developed under NSF grant IIS-1162046.

7. REFERENCES

[1] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. IEEE ICASSP, 1992*, pp. 517-520. Revision, February 19, 1997.

[2] H. Leung, and V. W. Zue, "A procedure for automatic alignment of phonetic transcription with continuous speech," *Proc. IEEE ICASSP 1984*, pp. 73-76, 1984.

[3] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication,* 12, pp. 357–370, 1993.

[4] A. Ljolje, J. Hirschberg, and J. van Santen, "Automatic speech segmentation for concatenative inventory selection," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 305-311, 1997.

[5] C. Wightman and D. Talkin, "The Aligner: Text to speech alignment using Markov Models," in J. van Santen, R. Sproat, J. Olive and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 313-323, 1997.

[6] J. P. Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, 2000.

[7] D. T. Toledano, L. A. H. Gomez. and L. V. Grande, "Automatic phoneme segmentation," *IEEE Trans. Speech and Audio Proc*., 11, pp. 617-625, 2003.

[8] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Proceedings of Acoustics*, pp. 5687-5690, 2008.

[9] K. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, 111, pp. 1872-1891, 2002.

[10] Y.-J. Kim and A. Conkie, "Automatic segmentation combining an HMM-based approach and spectral boundary correction," *Proc. ICSLP 2002*, pp.145-148, 2002.

[11] H.-Y. Lo and H.-M. Wang, "Phonetic boundary refinement using support vector machine," *Proc. IEEE ICASSP 2007*, pp. 933-936, 2007.

[12] D. T. Toledano, "Neural network boundary refining for automatic speech segmentation," *Proc.IEEE ICASSP 2000*, pp.3438-3441, 2000.

[13] K.-S. Lee, "MLP-based phone boundary refining for a TTS database," *IEEE Trans. Audio, Speech, and Language Proc.*, 14, pp. 981-989, 2006.

[14] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, "Phoneme alignment based on discriminative learning." *Proc. Interspeech*, pp. 2961–2964, 2005.

[15] J. P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, 51, pp. 352-368, 2009.

[16] L. Lehiste, "An acoustic-phonetic study of internal open juncture," *Phonetica*, 5, supplement, pp. 5-54, 1960.

[17] A. E. Turk and S. Shattuck-Hufnagel, "Word-boundary-related duration patterns in English," *Journal of Phonetics*, 28, pp. 397-440, 2000.

[18] M. Garellek, "Word-initial glottalization and voice quality strengthening," *UCLA Working Papers in Phonetics*, 111, pp. 92-122, 2012.

[19] L. H. Nakatani and K. D. Dukes, "Locus of segmental cues for word juncture," *J. Acoust. Soc. Am.*, 62, pp. 714-719, 1977.

[20] E. K. Johnson and P. W. Jusczyk, "Word segmentation by 8-month-olds: when speech cues count more than statistics," *Journal of Memory and Language*, 44, pp. 548-567, 2001.

[21] M. D. Tyler and A. Cutler, "Cross-language differences in cue use for speech segmentation," *J. Acoust. Soc. Am.*, 126, pp. 367-376, 2009.

[22] J. S. Garofolo, *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (LDC93S1), Linguistic Data Consortium, 1993.

[23] V. W. Zue and S. Seneff, "Transcription and alignment of the TIMIT database," in H. Fujisaki (ed.), *Recent Research Towards Advances Man-Machine Interface*, pp. 515-525, 1996.

[24] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech*," J. Acoust. Soc. Am.*, 87, pp. 1738-1752, 1990.

[25] The Hidden Markov Model Toolkit (HTK): http://htk.eng.cam.ac.uk/.

[26] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic Phonetic Segmentation using Boundary Models", *Proc. Interspeech*, 2013.

[27] V. R. R. Gadde, A. Stolcke, D. Vergyri, J. Zheng, K. Sonmez, and A. Venkataraman, "Building an ASR System for Noisy Environments: SRI's 2001 SPINE Evaluation System," *Proc. ICSLP*, vol. 3, pp. 1577-1580, Denver, 2002.

[28] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech and Language* 24(2), 273-288, 2010.

[29] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Proc*., vol.2, pp.578–589, 1994.

[30] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition", *Proc. IEEE ICASSP*, pp. 4117-4120, 2012.

[31] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, 1993.

[32] H. Teager, "Some observations on oral air flow during phonation", *IEEE Trans. Acoustic, Speech and Signal Proc.*, pp. 599–601, 1980.

[33] V. Mitra, M. McLaren, H. Franco, M. Graciarena, and N. Scheffer, "Modulation Features for Noise Robust Speaker Identification," *Proc. Interspeech*, pp. 3703-3707, 2013.

[34] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, "Medium duration modulation cepstral feature for robust speech recognition", *Proc. IEEE ICASSP*, Florence, 2014

[35] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients for robust speech recognition," *Proc. IEEE ICASSP*, pp. 4101-4104, 2012.

[36] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin, and J. H. L. Hansen, "Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA", *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, Atlanta, GA, USA, 2011.

[37] V. Mitra, H. Franco, and M. Graciarena, "Damped Oscillator Cepstral Coefficients for Robust Speech Recognition," *Proc. Interspeech*, pp. 886-890, 2013.

[38] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Juan, and M. Liberman, Articulatory features for large vocabulary speech recognition, *Proc. IEEE ICASSP*, pp. 7145-7149, 2013.

[39] G. Jekabsons, "M5PrimeLab: M5' regression tree and model tree toolbox for Matlab/Octave", 2010, available from http://www.cs.rtu.lv/jekabsons/.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML* pp. 807-814, 2010.

[41] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.