# FUSION OF ACOUSTIC, PERCEPTUAL AND PRODUCTION FEATURES FOR ROBUST SPEECH RECOGNITION IN HIGHLY NON-STATIONARY NOISE

*Ganesh Sivaraman[1], Vikramjit Mitra[2], Carol Y. Espy-Wilson[1]*

[1]University of Maryland College Park, MD USA
[2]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[1]{ganesa90, espy}@umd.edu, [2]vmitra@speech.sri.com

## ABSTRACT

Improving the robustness of speech recognition systems to cope with adverse background noise is a challenging research topic. Extraction of noise robust acoustic features is one of the prominent methods used for incorporating robustness in speech recognition systems. Prior studies have proposed several perceptually motivated noise robust acoustic features, and the normalized modulation cepstral coefficient (NMCC) is one such feature which uses amplitude modulation estimates to create cepstrum-like parameters. Studies have shown that articulatory features in combination with traditional mel-cepstral features help to improve robustness of speech recognition systems in noisy conditions. This paper shows that fusion of multiple noise robust feature streams motivated by speech production and perception theories help to significantly improve the robustness of traditional speech recognition systems. Key-word recognition accuracies on the CHiME-2 noisy-training task reveal that utilizing an optimal combination of noise robust features help to improve the accuracies by more than 6% absolute across all the different signal-to-noise ratios.

***Index Terms***— *Robust speech recognition, Modulation features, Articulatory features, Noise robust speech processing, Robust acoustic features, key word recognition.*

## 1. INTRODUCTION

Speech recognition in the presence of highly non-stationary noise is a challenging problem. There are many approaches that incorporate noise-robustness to automatic speech recognition (ASR) systems, including those based on 1) the feature space 2) the model space, and 3) missing feature theory. The approaches based on the model space and the marginalization based missing feature theories add robustness by adapting the acoustic model to reduce the mismatch between training and testing conditions. The feature-space approaches achieve the same by generating cleaner features for the acoustic model. Feature-space approaches can be classified into two subcategories. In the first subcategory, the speech signal is cleaned by using speech enhancement algorithms. (e.g., spectral subtraction, computational auditory scene analysis etc.). In the second subcategory, noise-robust acoustic features are extracted from the speech signal and used as input to the ASR system. Some well-known noise-robust features include power normalized cepstral coefficients (PNCCs) [1], fepstrum features [2] and perceptually motivated minimum variance distortion-less response (PMVDR) features [3]. Previous studies [4] have also revealed that articulatory features when used in combination with traditional acoustic features (e.g., mel-frequency cepstral coefficients or MFCCs) improve recognition accuracy of ASR systems.

In this paper we combine traditional cepstral features, perceptually motivated robust acoustic features and production-motivated articulatory features. The extracted features were deployed in the baseline small-vocabulary ASR system provided by the 2nd CHiME Challenge [5]. For our experiments we extracted a perceptually motivated feature: the Normalized Modulation Cepstral Coefficient (NMCC) [6] that analyzes speech using its estimated sub-band amplitude modulations (AMs). A detailed explanation of the NMCC feature is given in section 2. In addition, to the NMCC features, we explore the Vocal Tract constriction Variable (TV) trajectories [4] extracted from speech using a pre-trained artificial neural network. The estimated TVs have demonstrated significant noise robustness when used in combination with traditional cepstral features [4]. A detailed description of the TVs is given in section 3. Apart from these features, we have also used the traditional MFCCs (13 coefficients) along with their velocity ($\Delta$s), acceleration ($\Delta^2$s) and jerk ($\Delta^3$s) coefficients resulting in a 52D feature set.

The results obtained from different combinations of NMCC, TV and MFCC features show that the fusion of all the features provides better recognition accuracies compared to each individual feature. Section 4 describes the different combination of features that we have explored in our experiments.

The baseline system provided by the 2nd CHiME Challenge [5] was used as the speech recognizer. We also experimented with the parameters of the hidden Markov model (HMM) to arrive at the best configuration for our features. We present the model tuning steps in section 5 of the paper. The accuracy of the recognition results for the various features extracted are presented in section 6.

## 2. NMCC FEATURES

The Normalized Modulation Cepstral Coefficient (NMCC) [6] is motivated by studies [7, 8] showing that amplitude modulation (AM) of the speech signal plays an important role in speech perception and recognition. NMCC uses the nonlinear Teager's Energy Operator (TEO), Ψ, [9, 10], which assumes that a signal's energy is not only a function of its amplitude, but also of its frequency. Considering a discrete sinusoid $x[n]$, with $A$ = constant amplitude, $\Omega$ = digital frequency, $f$ = frequency of oscillation in hertz, $f_s$ = sampling frequency in hertz and $\theta$ = initial phase angle:

$$x[n] = A\cos[\Omega n + \theta];\ \Omega = 2\pi\,(f/f_s) \qquad (1)$$

If $\Omega \leq \pi/4$ and is sufficiently small, then Ψ takes the form

$$\Psi\{x[n]\} = \{x^2[n] - x[n-1]x[n+1]\} \approx A^2\Omega^2 \quad (2)$$

where the maximum energy estimation error in Ψ will be 23% if $\Omega \leq \pi/4$, or $f/f_s \leq 1/8$. The study discussed in [11] used Ψ to formulate the discrete energy separation algorithm (DESA), and showed that it can instantaneously separate the AM/FM components of a narrow-band signal using

$$\Omega_i[n] \approx \cos^{-1}\left\{1 - \frac{\Psi(x[n]) + \Psi(x[n+1])}{4\Psi(x[n])}\right\} \qquad (3)$$

$$|a_i[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1 - [\cos(\Omega_i[n])]^2}} \qquad (4)$$

Where $\Omega_i[n]$ and $|a_i[n]|$ denote the instantaneous FM signal and the AM signal, respectively, in the $i^{th}$ channel of the gammatone filterbank. Note that in (2) $x^2[n] - x[n-1]x[n+1]$ can be less than zero if $x^2[n] < x[n-1]x[n+1]$, while $A^2\Omega^2$ is strictly non-negative. In [6], we proposed to modify (2) into

$$\Psi\{x[n]\} = |\{x^2[n] - x[n-1]x[n+1]\}| \approx A^2\Omega^2 \quad (5)$$

which now tracks the magnitude of energy changes. Also, the AM/FM signals computed from (3) and (4) may contain discontinuities [12] (that substantially increase their dynamic range), for which median filters have been used. In order to remove such artifacts from the DESA algorithm, a modification was proposed in the AM estimation step in [6] followed by low-pass filtering.

The steps involved in obtaining the NMCC features are shown in Fig. 1. At the onset, the speech signal is pre-emphasized (using a coefficient of 0.97) and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. The windowed speech signal $s^w[n]$ is passed through a gammatone filterbank (using the configuration specified in [13].) with 50 channels spaced equally between 200 Hz to 7000 Hz in the ERB scale. The AM time signals $a_{k,j}[n]$ are then obtained for each of the 50 channels, where the total AM power of the windowed time signal for the $k^{th}$ channel and the $j^{th}$ frame is given as

$$P_{k,j}^{AM} = a_{k,j}^T a_{k,j} \qquad (6)$$

The resulting AM power is then power normalized, bias subtracted (as explained in [6]) and then compressed using the $1/15^{th}$ root, followed by the Discrete Cosine Transform (DCT) from which only the first 13 coefficients (including $C_0$) were retained. These 13 coefficients along with their Δs, $\Delta^2$s and $\Delta^3$s resulted in a 52D NMCC feature set.

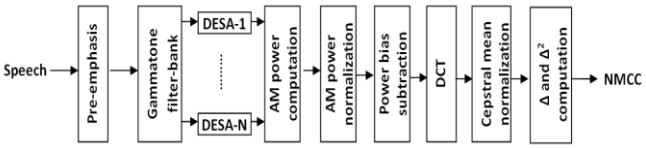


Figure 1: Flow-diagram of NMCC feature extraction from speech.

## 3. ARTICULATORY FEATURES

Previous studies [9, 10] have demonstrated that Artificial Neural Networks (ANNs) can be used to reliably estimate vocal tract constriction variable (Tract Variables also known as TV) trajectories [14] from the speech signal. TVs (refer to [14] for more details) are continuous time functions that specify the shape of the vocal tract in terms of constriction degree and location of the constrictors. Once trained, ANNs require low computational resources compared to other methods in terms of both memory requirements and execution speed.

An ANN has the advantage that it can have *M* inputs and *N* outputs; hence, a complex mapping of *M* vectors into *N* different functions can be achieved. In such architecture, the same hidden layers are shared by all *N* outputs, endowing the ANN with the implicit capability to exploit any correlation that the *N* outputs may have amongst themselves. The feed-forward ANN used in our study to estimate the TVs from speech were trained with back propagation using a scaled conjugate gradient (SCG) algorithm. To train the ANN model for estimating TVs, we need a speech database containing ground truth TVs. Unfortunately, since no such database is available at present, we used Haskins Laboratories' Task Dynamic model, (popularly known as TADA [17]) along with HLSyn [18] to generate a database containing synthetic speech along with articulatory specifications. From the CMU dictionary [19] 111,929 words were selected and their Arpabet pronunciations were input to TADA, which generated their corresponding TVs (refer to Table 1) and synthetic speech. Eighty percent of the data was used as the training set, 10% was used as the development set, and the remaining 10% was used as the test set. Note that TADA generated speech signals at a sampling rate of 8 kHz and TVs at a sampling rate of 200 Hz.

The input to the ANN was the speech signal parameterized as Normalized Modulation Cepstral Coefficients (NMCCs) [1], where 13 cepstral coefficients were extracted (note that the deltas were not generated from

these 13 coefficients) using a Hamming analysis window of 20 ms with a frame rate of 10 ms. These NMCC's are used as input features to the ANN model for estimating the TVs. They are different from the ones used for speech recognition given a different analysis window used. Note that telephone bandwidth speech was considered, where 34 gammatone filters spanning equally between 200 Hz to 3750 Hz in the ERB scale was used to analyze the speech signal. The TVs were downsampled to 100 Hz to temporally synchronize them with the NMCCs. The NMCCs and TVs were Z-normalized and scaled to fit their dynamic ranges into [-0.97, +0.97]. It has been observed [15] that incorporating dynamic information helps to improve the speech-inversion performance. In this case, the input features were contextualized by concatenating every other feature frame within a 200 ms window. Dimensionality reduction was performed on each feature dimension by using the DCT and retaining the first 70% of the coefficients, resulting in a final feature dimension of 104. Hence, for the TV estimator, *M* was 104 and *N* was 8 for the eight TV trajectories.

Initial experiments revealed that using temporally contextualized TVs as features provided better ASR performance than using the instantaneous TVs, indicating that the dynamic information of the TVs contributes to improving ASR performance. A context of 13 frames i.e., ~120 ms of temporal information was used to contextualize the TVs. To reduce the dimension of the contextualized TVs, the DCT was performed on each of the eight TV dimensions and their first seven coefficients were retained, resulting in a 56D feature set. We name this feature the modulation of TVs (ModTVs) [16].

## 4. FEATURE COMBINATIONS

The MFCCs used in all our experiments (except the baseline system, which used the HTK implementation of MFCCs [HTK-MFCC]) were obtained from SRI's Decipher® front end. Various combinations of the 52D MFCCs, 52D NMCCs and 56D ModTV features were experimented with. First, the MFCCs were combined with ModTVs to produce a 108 dimensional feature set. Then the dimensionality of the resulting feature was reduced to 42 for the noisy training setup using principal component analysis (PCA). The PCA transformation matrix was created such that more than 90% of the information is retained within the transformed features.

The PCA transformation matrix was learned using the training data and note that as per the 2nd CHiME challenge rules we have not exploited the fact that the same utterances were used within the clean and noisy training sets. These features were named as the MFCC+ModTV_pca. We also combined the 56D ModTV features with the 52D NMCC features and performed PCA on top of it and named it as NMCC+ModTV_pca, but the results from this experiment didn't show any improvement in recognition accuracy over the MFCC+ModTV combination.

We then explored a 3-way combination of NMCC, MFCC and ModTV features followed by PCA transform, that yielded 60D NMCC+MFCC+ModTV_pca feature. Note that in this case we observed that up to 60 dimensions after doing PCA transform retained more than 90% of the information.

Finally, we explored a combination of NMCC, MFCC and ModTV with utterance level mean and variance normalization that resulted in a 124D feature set after PCA transformation. In this case we noticed that 124 dimensions retained 90% of the information for the training datasets after PCA transformation. We name this feature as NMCC+MFCC+ModTV _mvn_pca. Figure 2 shows a block diagram representing all the feature combinations. The results obtained using these combination features is given in Table 1.

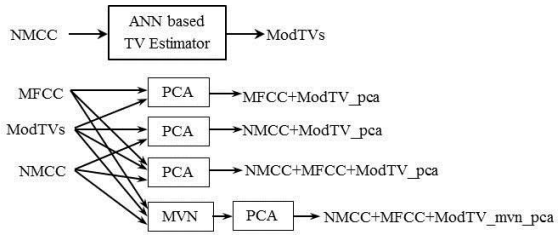


Figure 2: Block diagram showing the feature combinations

## 5. EXPERIMENTS AND RESULTS

### 5.1. Experiment settings

The data used in our experiments were obtained through the Track 1 of the 2nd CHiME Challenge. The dataset contained reverberated utterances recorded at 16 kHz sampling rate mixed with highly non-stationary background noise as described in [5]. The utterances consist of 34 speakers reading simple 6-word sequences of the form <command:4><color:4><preposition:4><letter:25><number :10><adverb:4>, where the numbers in brackets indicate the number of choices at each point [5]. The letters and numbers are the keywords in the utterances and the performance of the system was evaluated based on the recognition accuracy of these keywords.

We explored different features and their combinations as input to the whole-word small vocabulary ASR system distributed with the 2nd CHiME Challenge [5]. The baseline system used 39D MFCCs (after cepstral mean removal) obtained from HTK frontend [5]. The baseline recognizer uses whole word left-to-right hidden Markov models (HMMs) containing 51 words. The HMMs allowed no skips over the states and used 7 Gaussian mixtures per state with diagonal covariance matrices. The number of states for each word was based on 2 states per phoneme assumption and more details on the system topology are provided in [5].

Since the dimensionality of our input features varied from that used in the baseline system, we tuned the system configuration using the development set, by changing the number of states per phoneme, number of Gaussians per state, and the number of iterations for HMM parameter re-

estimation. The number of Gaussians was varied from 2 to 13. The number of iterations was varied from 4 to 8.

### 5.2. Results for Development set

We performed experiments on the development set in a systematic fashion in order to discover the best performance of the different feature sets. First, we conducted experiments using the baseline system provided with the 2nd CHiME challenge [5]. The keyword recognition accuracy results obtained for all the features from this experiment are provided in Table 1. After identifying the best feature sets, we tuned the system by varying the number of Gaussians from 2 to 13. Using the best tuned models for each feature set, we evaluated the test set results.

Initially, we tried the individual features: ModTVs (56D), MFCC (52D) and NMCC (52D) as input to the baseline HMM recognition system and observed that the NMCC feature provided the most improvement in recognition accuracy followed by the MFCC (52D) feature set. We also observed that the ModTVs by themselves were not showing any improvement in recognition accuracies over the baseline. The NMCC features by themselves demonstrated on an average 1.36% absolute improvement of the key word recognition accuracy over the baseline system.

As a next step we tried 2-way fusion, where we explored the following feature combinations: (1) MFCC+ModTV and (2) NMCC+ModTV. Both of these combinations yielded 108D features but they were reduced to 42D using PCA as discussed before. From these experiments we observed that adding the ModTVs to the MFCCs showed substantial improvement in performance, where the recognition accuracies were even better than the individual NMCC system. Unfortunately, the ModTVs didn't fuse well with the NMCCs. This might be because ModTVs were extracted using NMCCs instead of MFCCs as input to the ANN model as shown in Figure 2. We believe that the MFCC-ModTV fusion benefited from the amount of complimentary information they capture, whereas the TVs in reality being a non-linear transformation of NMCCs did not posses much complementary information compared to the NMCCs; hence their fusion (NMCC+ModTVs) did not do so well compared to the individual NMCC system.

As a final step, we fused the three features: NMCC, ModTVs and MFCCs together and performed PCA on top of it to produce a 60D feature set and this fusion gave an average improvement of around 1.45% absolute over the baseline system. This showed that even though NMCCs by themselves didn't fuse so well with the ModTVs, a 3-way combination yielded the best recognition accuracy compared to the individual feature based systems and 2-way fusion based systems.

Note that we did not implement any utterance-level mean and variance normalization across all of the feature dimensions in any of the fusion strategies discussed above. Hence to observe if such normalization helps to further improve the recognition accuracies, we remade the 3-way combination followed by utterance level mean and variance normalization followed by PCA transform. At this step we observed that 90% of the information resided in the top 124 dimensions, hence we generated a 124D feature set from this mean-variance normalized 3-way fused feature set. Results on the development set showed an average 2.17% absolute improvement of the recognition accuracies over the baseline.

After evaluating the feature sets on the baseline system, we selected the best performing features namely – NMCC, MFCC+ModTV_pca, NMCC+MFCC+ModTV_pca and NMCC+MFCC+ModTV_mvn_pca. We then tuned the models for each of these feature set by varying the number of Gaussians from 2 to 13 and the number of parameter re-estimation iterations from 4 to 8. The results obtained by varying the number of Gaussians in the mixture for the NMCC+MFCC+ModTV_pca feature are shown in table 3. The keyword recognition accuracies using the tuned models for the development sets of the selected features are shown in Table 3. Note that the tuned parameters for each of the features presented in Tables 3 and 4 are not the same. However for the sake of brevity we are providing the parameters for only the best system. For the others, the tuned parameters were very similar (if not same) to the best system.

**Table 1**: *Keyword recognition accuracy in percent for the development set with noisy trained models using the baseline system having 7 Gaussian mixtures per state*.

| Features | -6dB | -3dB | 0dB | 3dB | 6dB | 9dB | Average |
|---|---|---|---|---|---|---|---|
| Baseline MFCC (39D) [HTK-MFCC] | 49.67 | 57.92 | 67.83 | 73.67 | 80.75 | 82.67 | 68.75 |
| MFCC (52D) | 47.50 | 55.00 | 65.92 | 73.08 | 79.58 | 82.75 | 67.31 |
| ModTV (56D) | 16.00 | 19.42 | 20.83 | 25.33 | 30.50 | 32.50 | 24.09 |
| NMCC (52D) | 51.17 | 59.25 | 68.58 | 75.42 | 81.75 | 84.50 | 70.11 |
| MFCC+ModTV_pca (42D) | 51.42 | 60.50 | 69.83 | 76.08 | 81.08 | 84.83 | 70.62 |
| NMCC+ModTV_pca (42D) | 43.92 | 52.92 | 59.58 | 69.92 | 75.83 | 77.67 | 63.30 |
| NMCC+MFCC+ModTV_pca (60D) | 50.58 | 60.08 | 69.00 | 76.75 | 81.08 | 83.75 | 70.20 |
| NMCC+MFCC+ModTV_mvn_pca (124D) | 53.33 | 60.92 | 67.33 | 75.67 | 82.50 | 84.00 | 70.62 |

**Table 2**: *Keyword recognition accuracy in percent for the development set with noisy trained models by tuning the number of Gaussians per mixture per state. [Results are for* NMCC+MFCC+ModTV_pca *feature set]*

| Number of Gaussians | -6dB | -3dB | 0dB | 3dB | 6dB | 9dB | Average |
|---|---|---|---|---|---|---|---|
| 2 | 53.08 | 62.58 | 72.33 | 78.67 | 85.50 | 86.25 | 73.06 |
| 3 | 53.92 | 64.25 | 73.00 | 79.83 | 86.42 | 87.17 | 74.09 |
| 5 | 53.00 | 62.50 | 71.25 | 78.08 | 82.75 | 85.00 | 72.09 |
| 7 | 50.58 | 60.08 | 69.00 | 76.75 | 81.08 | 83.75 | 70.20 |
| 11 | 47.58 | 57.00 | 66.08 | 73.67 | 78.17 | 80.83 | 67.22 |
| 13 | 45.92 | 55.33 | 64.75 | 71.75 | 76.92 | 80.58 | 65.87 |

**Table 3**: *Keyword recognition accuracy in percent for the development set with noisy trained models after tuning*

| Features | -6dB | -3dB | 0dB | 3dB | 6dB | 9dB | Average |
|---|---|---|---|---|---|---|---|
| Baseline MFCC (39D) [HTK-MFCC] | 49.67 | 57.92 | 67.83 | 73.67 | 80.75 | 82.67 | 68.75 |
| NMCC (52D) | 56.08 | 62.75 | 71.5 | 78.5 | 84.17 | 86.58 | 73.26 |
| MFCC+ModTV_pca (42D) | 53.17 | 63.00 | 71.00 | 79.92 | 85.25 | 87.33 | 73.27 |
| NMCC+MFCC+ModTV_pca (60D) | 53.92 | 64.25 | 73.00 | 79.83 | 86.42 | 87.17 | 74.09 |
| **NMCC+MFCC+ModTV_mvn_pca (124D)** | **58.08** | **65.00** | **74.42** | **81.00** | **85.58** | **86.58** | **75.11** |

**Table 4**: *Keyword recognition accuracy in percent for the test set with noisy trained models after tuning*

| Features | -6dB | -3dB | 0dB | 3dB | 6dB | 9dB | Average |
|---|---|---|---|---|---|---|---|
| Baseline MFCC (39D) [HTK-MFCC] | 49.33 | 58.67 | 67.50 | 75.08 | 78.83 | 82.92 | 68.72 |
| NMCC (52D) | 54.42 | 64.58 | 73.08 | 80.08 | 82.92 | 87.92 | 73.83 |
| MFCC+ModTV_pca (42D) | 53.67 | 63.67 | 72.50 | 79.83 | 84.00 | 87.33 | 73.50 |
| NMCC+MFCC+ModTV_pca (60D) | 55.75 | 64.08 | 73.17 | 81.25 | 84.83 | 87.83 | 74.48 |
| **NMCC+MFCC+ModTV_mvn_pca (124D)** | **57.75** | **65.42** | **75.08** | **82.25** | **85.75** | **87.33** | **75.59** |

### 5.3. Results for the Test set

Using the models tuned on the development set for each feature, we evaluated the corresponding feature's test set results. Table 4 shows the keyword recognition accuracy for the test set using the tuned acoustic models trained with noisy speech data. The NMCC feature gave an average of 5% absolute improvement in accuracy over the baseline. The MFCC+ModTVs_pca feature provided an average of 5% absolute improvement in accuracy over the baseline, indicating that the acoustic models trained with NMCC and MFCC+ModTV had similar performance. The NMCC + MFCC + ModTVs_pca feature gave an average of 6% absolute improvement over the baseline, indicating that the three way feature combination offered the best performance. Finally, the mean and variance normalized features NMCC+MFCC+ModTVs_mva_pca provided an average of 7% absolute improvement in keyword recognition accuracy over the baseline and this setup gave the best performing results from our experiments for the 2nd CHiME challenge.

## 6. CONCLUSIONS

Our experiments presented a unique combination of traditional acoustic features, perceptually motivated noise robust features and speech production based features and showed their combination gave the best keyword recognition accuracy compared to their individual performance. NMCC was found to be the best performing single feature for the given key-word recognition task and its performance was further improved when combined with the MFCCs and ModTVs. The success in the 3-way combination of the features lies in their mutual complementary information. Our experiments mostly focused on the front-end feature exploration with no alteration of the backend recognizer, except HMM parameter tuning. In the future, we want to explore enhanced acoustic modeling schemes which can further improve the recognition accuracies. Many researchers have hypothesized that the combination of perceptual, production and acoustic features will result in a superior front end for

speech recognition systems. The experiments presented here support this hypothesis with data.

## 7. ACKNOWLEDGEMENT

This research was supported by NSF Grant # IIS-1162046.

## 8. REFERENCES

[1] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," *in Proc. of ICASSP*, pp. 4574–4577, 2010.

[2] V. Tyagi, "Fepstrum features: Design and application to conversational speech recognition," *IBM Research Report*, 11009, 2011.

[3] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Comm.*, vol.50, iss.2, pp. 142–152, 2008.

[4] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, "Tract variables for noise robust speech recognition," *IEEE Trans. on Audio, Speech & Language Processing*, 19(7), pp. 1913-1924, 2011.

[5] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta and Marco Matassoni, "The Second `CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," *in Proc. of ICASSP*, May 26-31, 2013.

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

[7] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. of Am.*, 95(5), pp. 2670–2680, 1994.

[8] O. Ghitza, "On the upper cutoff frequency of auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. of Am.*, 110(3), pp. 1628–1640, 2001.

[9] H. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. ASSP*, pp. 599–601, 1980.

[10] J.F. Kaiser, "Some useful properties of the Teager's energy operator," *in Proc. of IEEE*, Iss. III, pp. 149-152, 1993.

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

[12] J.H.L. Hansen, L. Gavidia-Ceballos, and J.F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomedical Engineering*, 45(3), pp. 300–313, 1998.

[13] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, 47, pp.103–138, 1990.

[15] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. 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.

[16] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan and M. Liberman, "Articulatory trajectories for large-vocabulary speech recognition," *to appear*, ICASSP 2013.

[17] H. Nam, L. Goldstein, E. Saltzman and D. Byrd, "Tada: An enhanced, portable task dynamics model in Matlab," *J. of Acoust. Soc. Am.*, 115(5), pp. 2430, 2004.

[18] H. M. Hanson and K. N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," *J. of Acoust. Soc. Am.*, 112(3), pp. 1158-1182, 2002.

[19] http://www.speech.cs.cmu.edu/cgi-bin/cmudict