Temporal Modulation Processing of Speech Signals for Noise Robust ASR

Hong You, Abeer Alwan

Department of Electrical Engineering, University of California, Los Angeles, CA 90095

hyou@ee.ucla.edu, alwan@ee.ucla.edu

Abstract

In this paper, we analyze the temporal modulation characteristics of speech and noise from a speech/non-speech discrimination point of view. Although previous psychoacoustic studies [3][10] have shown that low temporal modulation components are important for speech intelligibility, there is no reported analysis on modulation components from the point of view of speech/noise discrimination. Our data-driven analysis of modulation components of speech and noise reveals that speech and noise is more accurately classified by low-passed modulation frequencies than band-passed ones. Effects of additive noise on the modulation characteristics of speech signals are also analyzed. Based on the analysis, we propose a frequency adaptive modulation processing algorithm for a noise robust ASR task. The algorithm is based on speech channel classification and modulation pattern denoising. Speech recognition experiments are performed to compare the proposed algorithm with other noise robust frontends, including RASTA and ETSI AFE. Recognition results show that the frequency adaptive modulation processing is promising.

Index Terms: noise robust speech recognition, temporal modulation processing, spectro-temporal processing

1. Introduction

In this study, we aim to enhance the noise robustness of an automatic speech recognition (ASR) system via a front-end approach. The focus is to incorporate adaptive amplitude modulation processing into the standard short-time based feature extraction algorithm. As a result, ASR noise robustness is improved.

Studies in speech perception have confirmed the importance of amplitude modulation frequencies on speech intelligibility. Drullman et. al [10] shows that modulation frequencies below 16Hz contribute to speech intelligibility significantly, especially for vowel perception in noise. In [3], a quantitative model of amplitude modulation is described that explains data from modulation detection and modulation masking experiments.

There is also increasing interest in using modulation information for ASR. The relative spectra (RASTA) algorithm[11] filters the temporal envelope trajectory to minimize convolutional channel effects on speech signals and to emphasize low modulation frequencies. Using RASTA, ASR under both additive and convolutional noise is considerably improved. In [4], temporal information is studied in an effort to search for speech features invariant to both noise and speaker differences. These are reached by emphasizing temporal information around 4 Hz, and attenuating one that moves beyond a syllable rate (2-12Hz). The stability of the representation against additive noise is demonstrated by spectrogram-like displays.

Recent work on modulation frequencies are further influenced by current psychoacoustic studies of modulation information, and a computational spectro-temporal model developed in [5].

There are two main types of algorithms incorporating modulation frequencies. One is based on joint spectro-temporal features, and the other is based on sequential spectrum and temporal processing. Spectro-temporal processing, as in [1] and [2], is developed to denoise speech signals according to spectro-temporal SNR estimation. High dimensionality is a common problem for features in the joint spectro-temporal domain. Although some solutions have been developed ([1] and [9]), dimensionality reduction of spectro-temporal features remains a challenge. An example of the second type of methods (sequential spectrum processing and temporal processing) is a data-driven temporal feature processing algorithm (TRAP_TANDEM)[8] to further improve upon the original RASTA processing. Although some studies have explored effective information combination[6] for this sequential type of processing, ways to combine spectral information and temporal information need to be further investigated.

The present work addresses noise robustness using the sequential type of algorithms. However, it differs from previous approaches in several ways. First, we derive a frequency-adaptive modulation domain processing to denoise speech signals. We show that frequency adaptiveness can significantly improve temporal modulation processing performance. Second, characteristics of speech and noise in the modulation frequency (MF) are studied, as well as the effects of additive noise on modulation characteristics. In contrast to previous analysis that have focused on speech intelligibility, the focus of the present analysis is on speech and noise separability in MF. Furthermore, a denoising algorithm is developed to attenuate noise sensitive MFs based on estimates from noise robust MFs.

Although frequency channel SNR estimation using temporal MFs has been studied in [13], our noisiness indicator differs from that of previous work. Beside temporal scale differences, (600ms in our algorithm and around 40ms in [13]), we use a noisiness indicator based on the decrease in MFs between 0 Hz and 1.5 Hz, whereas SNR estimation in [13] is for short-time frequency denoising, relying more on pitch information derived from high modulation frequencies.

The remaining of this paper is organized as follows. Section 2 presents analysis of temporal modulation frequencies of both speech and noise signals, and compares different modulation features for a speech/non-speech discrimination task. Section 3 describes the proposed modulation processing algorithm. Speech recognition experiments are described in Section 4, followed by a summary in Section 5.

2. Analysis

2.1. Speech/Non-speech Classification Experiment

Because speech and noise signals have distinct modulation frequency characteristics, we test this feature separability using a classification experiment.

Modulation frequencies (from 0 Hz to 16 Hz) are used as temporal modulation features. The dimension of temporal modulation feature varies according to our experimental setting. For example, $0-2$ Hz feature in Figure 1 is of dimension 3, and $0-16$ Hz feature is of dimension 11. A long window length is used for computing modulation features (600ms). In addition, we aggregate speech and noise modulation features for training and testing according to acoustic frequencies. We perform a feature classification task using linear support vector machines (SVMs) trained to classify speech/noise modulation features.

Speech data from TIMIT are used; babble noise data are synthesized from randomly sampled TIMIT data so that the babble noise density is controlled. Overall, 10 minutes of speech and babble noise are used to train the classifier, and 10 minutes of speech and noise data are used in testing.

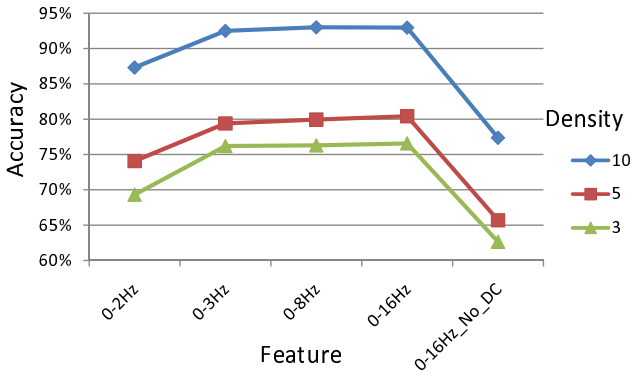

Figure 1: Classification accuracy for modulation features

Figure 1 depicts SVM classification results. Performance curves corresponding to classification of speech and babble noise of different densities (10, 5, and 3) are shown. The density parameter controls the number of random speech signals used in generating synthetic noise. Density 3 represents a more challenging task than density 10. As shown in the figure from left to right, tested modulation features included low-passed MFs, and a band-passed MF (the 0-16Hz-No-DC feature) where the 0 Hz MF is excluded. Band-passed MF performs consistently worse than low-passed ones. The results show the importance of combining DC modulation along with low modulation frequency for the task of speech/non-speech feature classification. Therefore, although the removal of long-term static magnitude spectrum negatively affects speech intelligibility to a small extent[10], the micro-structure of frequency energy distribution provided by the static magnitude spectrum appears to be important for speech/non-speech classification.

2.2. Speech and Noise Modulation Characteristics

Similar to studies on modulation characteristics of speech[4], we observe that speech MFs have smoother energy transition from 0 Hz to low modulations, and a localized MF peak at around 4 Hz. For noise, however, sharp energy decrease occur from 0 Hz to low MFs. This may be be a good feature for speech/non-speech discrimination in a frequency channel.

Unlike existing analysis that studies modulation characteristics for a given acoustic frequency, we analyze the cross correlation between different MFs across acoustic frequencies. A high cross correlation between different MFs indicates the existence of consistent modulation patterns across acoustic frequencies, whereas a low cross correlation indicates less consistent modulation patterns. Based on speech production theory, it is our hypothesis that the modulation pattern of speech signals over long segments is more consistent than that of noise.

Figure 2 presents the histogram of cross correlation between low MFs 1.5Hz and 3Hz. For speech signals, the histogram shows that low MFs are highly correlated, while cross correlations for noise signals are mostly less than 0.9 between low modulation frequencies.

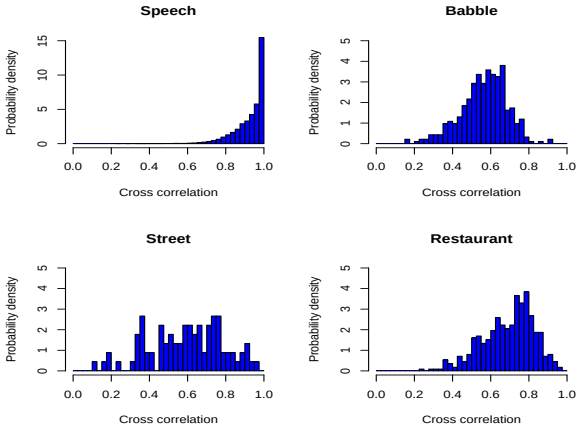

Figure 2: Histogram of cross correlation between MF 1.5Hz and 3Hz. The x-axis is cross correlation; The y-axis is probability density. Upper left and right panels are speech and babble noise; bottom left and right are street noise and restaurant noise.

Figure 3 presents the cross correlation histogram of speech signal corrupted by subway noise at SNR 15dB and 5dB. Additive noise considerably reduces cross correlation between low MFs. Therefore, a modulation processing algorithm needs to address this reduced cross-correlation between low MFs due to noise to avoid modulation pattern mismatch for noise robust ASR.

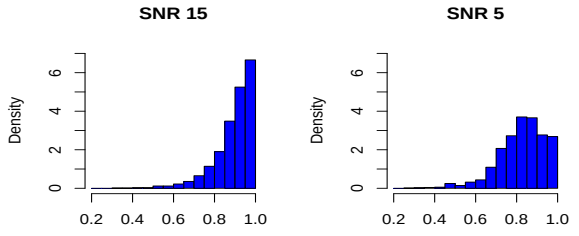

Figure 3: Cross correlation between modulation frequencies $1.5Hz$ and $3Hz$ of noisy speech at SNR 15dB and 5dB (left and right panels respectively).

3. Method

A frequency adaptive modulation processing algorithm is developed. The basic idea is to denoise noise-sensitive MFs for noisy channels based on a linear combination of noise-sensitive and noise-robust MFs from the the closet robust channel and noise-robust MFs from the noisy channel under consideration. We assume that for most noise types and SNRs, noise-sensitive MFs range from 0 Hz to 2 Hz, whereas noise robust MFs range from 3 Hz to 8 Hz[4].

In contrast to RASTA[11] in which the static average magnitude spectrum (i.e. the micro-structure information contained in modulation 0Hz) is filtered to remove convolutional noise, we adaptively denoise MFs less than 2 Hz. The benefits of recovering this micro-structure information have been discussed in [7]. The author suggests that this information is important for avoiding negative reconstructed magnitude spectrum. In addition, our speech/non-speech feature classification experiment shows modulation at 0 Hz is important when combined with low MFs in order to provide a salient speech event cue. Accordingly, it plays two distinct roles in our algorithm. First, when combined with the MF at 1.5Hz, it can be used as a frequency channel SNR correlated indicator or "noisiness" indicator. Second, denoised MFs at 0 Hz and low MFs can provide speech detection information for the front-end. Without the 0 Hz MF, speech detection in noise is harder, especially when negative magnitude spectrum issues occur. For that reason, modulation at 0 Hz is also denoised in our algorithm rather than attenuated.

The general modulation processing flow chart is shown in Figure 4.

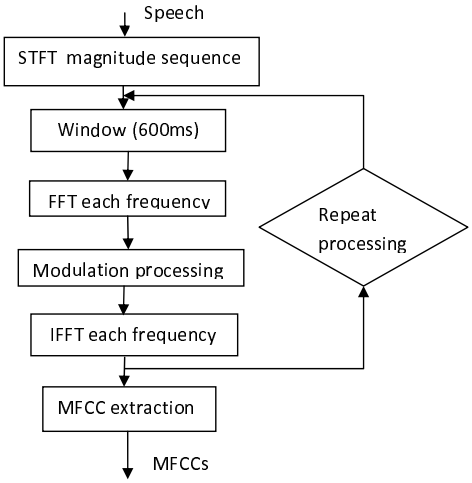

Figure 4: Flowchart of modulation domain processing

Implementation of the adaptive modulation processing is shown in Figure 5. $d_i(f)$ is the log magnitude at MF i and frequency f, while $d_i^{'}(f)$ corresponds to denoised log magnitude. A near log magnitude compression $20log_{10}(1+|x|)$ is used to reduce the dynamic range of modulation components. MFs $\leq$ 2 Hz are denoised, while MFs $>$ 16 Hz are attenuated. Each module of the algorithm in this figure is explained below in detail.

As shown, a modulation energy measure $\beta(f)$ and a noisiness indicator $\xi(f)$ are extracted. Specifically, $\beta(f) = \sum_{i=3,..,8} d_i(f)$, and $\xi(f) = d_0(f) - d_{1.5}(f)$. These two measurements are used to select acoustic frequencies with high

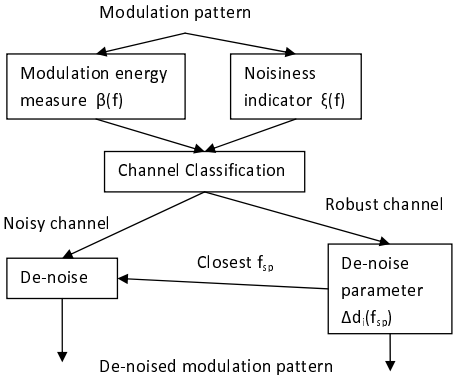

Figure 5: Flowchart of frequency adaptive modulation processing algorithm

modulation energy and small $\xi(f)$ by a channel classification module. Small $\xi(f)$ indicates a smoother energy transition from modulation 0 Hz to 1.5 Hz, thus a higher SNR for the channel. To classify frequency channels, a recursive min-max scheme is used. Noise robust channels are selected every 200Hz in our implementation. There are 2 thresholds used in the classification module: T_β and T_ξ. T_β is an adaptive threshold adjusted according to local maximum of $\beta(f)$, while T_ξ is an empirical threshold for $\xi(f)$. $T_\xi = 5.5 + 0.33\times$ mean of $\xi(f)$ over $0 < f < 50$. Frequencies corresponding to smaller $\beta(f)$ compared to T_β are removed. Among the remaining frequencies, frequency with minimum $\xi(f)$ is chosen as a possible noise robust channel. To reduce misclassification when SNRs are low, threshold T_ξ is used as a upper bound for the possible noise robust channel. The min-max computation is then repeated throughout the whole frequency range. This way, the channel classification module is optimized to minimize false classification errors.

Speech modulation parameters, $\Delta d_i(f_{sp})$ for $i \leq 2$ Hz, are subsequently estimated from the classified noise robust channel f_{sp}. Specifically, for $i = 0, 1, 2$ Hz,

$$\Delta d_i(f_{sp}) = d_i(f_{sp}) - \frac{1}{3}\Sigma_{j=3,4,5} d_j(f_{sp}). \quad (1)$$

When assuming an exponential model for MF magnitudes (approximately holds for speech MFs), $\Delta d_i(f_{sp})$ is close to the log linear parameter for the modulation magnitude at frequency f_{sp} and MF i. Therefore, for a noisy channel f, the closest f_{sp} is used to denoise noisy MFs. For $i \leq 2$ Hz,

$$d_i^{'}(f) = \Delta d_i(f_{sp}) + \frac{1}{3}\Sigma_{i=3,4,5} d_i(f) \quad (2)$$

Finally, the denoised MF $d_i^{'}(f)$ is compared with the corresponding noisy MF $d_i(f)$. A noisy MF is denoised if $|d_i^{'}(f) - d_i(f)| > K$dB, where K is 1.5dB in our implementation.

4. Experiments

A noise robust recognition experiment is set using Aurora2 database. HMM models are configured and trained according to standard Aurora2 setting using HTK tools. Word level acoustic models are trained, where each model is represented by 18 states and 3 mixtures per state. The MFCC features (plus first- and second- order derivatives) are used as our baseline system. We compare the ASR recognition accuracy of our

method, the frequency adaptive modulation processing, to that of MFCC, RASTA-MFCC, and ETSI advanced frontend[12] (ETSI-AFE). A RASTA-MFCC is experimented instead of the original RASTA-PLP because the performance of the former is significantly better than the latter on Aurora dataset. In addition, RASTA-MFCC provides a fair comparison between our algorithm and the original bandpass based modulation processing idea. Table 1 shows results of these front-ends tested on set A subway noise.

Table 1: *ASR accuracy of MFCC, RASTA-MFCC, ETSI-AFE, and the proposed algorithm (Adaptive) under subway noise in Aurora set A*

	MFCC	RASTA-MFCC	ETSI-AFE	Adaptive
SNR20	97.05	95.00	98.4	95.60
SNR15	93.49	90.94	96.32	94.23
SNR10	78.72	80.2	91.53	86.34
SNR5	52.2	63.83	77.92	70.34
SNR0	26.01	37.98	50.91	43.75
SNR-5	11.18	16.79	20.4	18.67
Average	59.77	64.13	72.58	68.16

The average performance (averaged over SNR from 20 dB to -5 dB) of set A and set B is shown in Table 2 and Table 3.

Table 2: *Average ASR accuracy (20 dB to -5 dB) of MFCC, RASTA-MFCC, ETSI-AFE, and the Adaptive algorithm of Aurora set A*

	MFCC	RASTA	ETSI-AFE	Adaptive
Subway	59.77	64.13	72.58	68.16
Babble	41.83	63.40	71.93	60.95
Car	52.06	63.06	73.75	71.86
Exhibition	56.09	63.68	73.09	68.21
Average	52.43	63.56	72.83	67.30

Table 3: *Average ASR accuracy (20 dB to -5 dB) of MFCC, RASTA-MFCC, ETSI-AFE, and the Adaptive algorithm of Aurora set B*

	MFCC	RASTA	ETSI-AFE	Adaptive
Restaurant	44.40	66.10	73.14	62.48
Street	53.01	64.21	72.42	69.24
Airport	45.74	70.12	75.65	72.36
Train	47.76	64.94	75.74	70.96
Average	47.72	66.34	74.23	68.76

The proposed frequency adaptive modulation algorithm improves noise robust ASR performance considerably over RASTA algorithm for all cases except for babble-like noise. This result indicates that a frequency adaptive denoising scheme in the modulation domain is preferable to a fixed filtering design. For babble noise, the proposed algorithm performs less well than the RASTA algorithm. This maybe is due to the fact that babble noise has smoother modulation energy distribution than other types of noise. Hence, it has more false classification errors in the channel classification modulation. Although ASR performance of our method is worse than that of the ETSI-AFE noise robust frontend, we believe that our algorithm remains promising as it uses only simple modulation domain processing, whereas ETSI-AFE utilizes a set of complex optimized noise robust techniques.

5. Conclusions

In this work, we study the modulation characteristics of speech and noise, and propose frequency selective modulation domain processing for noise robust ASR. We analyze the modulation domain from the point of view of speech and noise discrimination. Our data exploration indicates that denoising processing in the modulation domain is promising when a frequency adaptive scheme is applied. Key observations obtained in our study include: distinction of speech and noise in modulation characteristics; importance of micro-structure information for speech/non-speech classification; modulation frequency correlation over the acoustic frequencies of speech and noise. These observations motivate us to design a frequency adaptive modulation processing algorithm that improves noise robust ASR. In addition, our study indicates a need to adaptively denoise modulation frequencies of noisy speech, while utilizing noise robust modulation pattern detected across frequency channels.

6. References

[1] Nima Mesgarani, Malcolm Slaney, Shihab Shamma, "Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, 2006, page 920-930.

[2] Nima Mesgarani, Shihab Shamma, "Denoising in the Domain of SpectrotemporalModulations", EURASIP Journal on Audio, Speech, and Music Processing Volume 2007.

[3] Torsten Dau, Birger Kollmeier, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers", Journal Acoustical Society of America, Nov., 1997, page 2892-2905.

[4] Steven Greenberg, Brian Kingsbury, "The Modulation Spectrogram: In pursuit of an invariant representation of speech", ICASSP, 1997, page 1647-1650.

[5] Taishih Chi, Powen Ru, and Shihab A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds", Journal Acoustical Society of America, Aug., 2005, page 887-906.

[6] Fabio Valente and Hynek Hermansky, "On the Combination of Auditory and Modulation Frequency Channels for ASR applications", interspeech 2008, page. 2242-2245.

[7] T. H. Falk, S. Stadler, W.B. Kleijn, and Wai-Yip Chan, "Noise Suppression Based on Extending a Speech-Dominated Modulation Band", Interspeech 2007, page 970-973.

[8] Hynek Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech", ASRU 2003.

[9] Sherry Y Zhao, Nelson Morgan, "Multi-stream spectro-temporal features for robust speech recognition", interspeech 2008, page 898-901.

[10] Rob Drullman, Joost M. Festen, and Reinier Plomp, "Effect of temporal envelope smearing on speech reception", Journal acoustical society of america, Feburary 1994, page 1053-1064.

[11] Hynek Hermansky, Nelson Morgan, "RASTA Processing of Speech", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 2. NO. 4,OCT, 1994

[12] ETSI standard document-ETSI ES 202 050 v1.1.1, "Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression Algorithm", 2002.

[13] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with application to noise suppression", IEEE trans. Speech and Audio Proc., Vol. 11, No.3, May 2003.