Multistream Bandpass Modulation Features for Robust Speech Recognition

Sridhar Krishna Nemala, Kailash Patil, Mounya Elhilali

Department of Electrical and Computer Engineering
Center for Speech and Language Processing
The Johns Hopkins University Baltimore MD USA
{nemala,kailash,mounya}@jhu.edu

Abstract

Current understanding of speech processing in the brain suggests dual streams of processing of temporal and spectral information, whereby slow vs. fast modulations are analyzed along parallel paths that encode various scales of information in speech signals. This unique way for the biology to analyze the multiplicity of information in speech signals along parallel paths can bare great lessons for feature extraction front-ends in speech processing systems, particularly for dealing with extrinsic degradations and unseen noise distortions. Here, we propose a multistream approach to feature analysis for robust speaker-independent phoneme recognition in presence of nonstationary background noises. The scheme presented here centers around a multi-path bandpass modulation analysis of speech sounds with each stream covering an entire range of temporal and spectral modulations. By performing bandpass operations of slow vs. fast information along the spectral and temporal dimensions, the proposed scheme avoids the classic feature explosion problem of previous multistream approaches while maintaining the advantage of parallelism and localized feature analysis. The proposed architecture results in substantial improvements over standard baseline features and two state-of-the-art noise robust feature schemes.

Index Terms: multistream, specto-temporal modulations, speech recognition, noise robustness

1. Introduction

Speech operates on multiple time scales, ranging from a few milliseconds to hundreds of milliseconds; each with its distinct acoustic manifestation, neural instantiation and perceptual role. Classic paradigms for speech processing generally ignore the intricate interplay between these various time constants, and process information on one time scale (typically, segment by segment), only to integrate it at later stages in a feedforward fashion (via temporal derivatives and contextual information). In contrast, evidence from neurophysiology and neurolinguistic literature argues for at least a dual parallel processing mode [1]. In this view, a shorter time scale of the order of 15Hz (roughly 30-80msec) captures segmental transitions in speech; while a longer time scale of the order of 5Hz (roughly 150-300msec) is commensurate with the size of syllable transitions in speech. This multi-stream framework for speech processing has the advantage of capturing the multiple, partially redundant, cues in speech and benefiting from the parallel processing mode in order to achieve more stable recognition in presence of background noise and distortions. A similar framework is also advantageous for processing spectral cues in speech, which also contain separate yet somewhat redundant information about the phonetic identity of the signal [2, 3]. In this regard, one can capture the broad trends of the spectral shape (e.g. formants) distinctly from the rapidly varying properties (e.g. harmonic peaks). Put together, we view the dual-mode of processing slow vs. fast temporal and spectral cues as a propitious framework for parallel processing of speech information which could bear great benefits not only for improved automatic recognition, but more importantly robust performance in presence of degradations.

Here, we present a multistream framework for automatic speech recognition (ASR) that integrates multiple streams spanning slow vs. fast dynamics of speech, both spectrally and temporally. In this scheme, the multiple feature streams are constructed based on bandpass modulation filtering, with each stream covering a full range of either slow or fast spectral and temporal modulations. This approach contrasts directly with previous attempts at incorporating multistream processing in ASR systems; whereby conventional Gabor filters centered at a number of specific temporal or spectral modulation frequencies are used [4, 5, 6]. These past attempts at multistream processing have generally used the rationale of feature division, using an array of features *localized* around a set of modulations. Such architectures have the obvious drawback of dimensionality explosion of the feature space into several thousands of dimensions or equivalently in several tens of feature streams [4, 5, 6]. In contrast, our proposed bandpass scheme of dual slow/fast processing maintains the benefits of parallel processing without *any* dimensionality expansion in the feature extraction stage.

We evaluate the benefits of this multistream bandpass modulation feature scheme by comparing its performance with standard baseline ASR features, Mel-Frequency Cepstral Coefficients (MFCC), and two state-of-the-art noise robust feature schemes namely Mean-Variance ARMA (MVA) processing [7] and Advanced-ETSI noise-robust speech recognition front-end [8]. In particular, we focus on the robustness of these different feature representations as a function of noise distortions in a speaker-independent phoneme recognition task. The following section presents motivation and details of the multistream feature representation. The experimental setup for the recognition task and results are detailed in section 3. In section 4, we finish with a discussion of these results and potential improvements towards achieving further robustness to noise distortions.

2. Multistream modulation features

The parametrization of speech sounds is achieved through a multistage model that captures processing taking place at various stages along the auditory pathway from the periphery all the way to the primary auditory cortex (A1). The input acoustic signal is first processed through a pre-emphasis stage, im-

plemented as a first-order highpass filter with pre-emphasis coefficient 0.97. An early stage then maps the one-dimensional acoustic signal to an auditory time-frequency spectrographic representation detailed in [9]. The first step consists of cochlear-filtering, using a bank of 128 constant-Q ($Q = 4$), highly asymmetric, bandpass filters, equally spaced on a logarithmic frequency axis (24 filters/octave over a 5.3 octave range). A subsequent step performs a sharpening of the filterbank frequency selectivity (from $Q = 4$ to 12) via a first-difference over neighboring channels, followed by half-wave rectification and short-term integration over $10msec$ windows, and a cubic-root compression of the spectrogram. Finally, the number of frequency channels is decimated by a factor of 4, resulting in 32 frequency channels with a resolution of 6 channels/octave over 5.3 octaves.

A central stage further analyzes the auditory spectrogram to form multiple feature streams. Each individual feature stream is obtained by filtering the auditory spectrogram using a set of bandpass spectral and temporal modulation filters. The filtering is done in the Fourier domain of the modulation amplitude. For the spectral (or temporal) modulation filtering, first the Fourier transform of each spectral (or temporal) slice in the spectrogram is taken, then is multiplied by a bandpass modulation filter $H(w;[w_l,w_u])$ capturing entire modulation content within the specified range of w_l and w_u ($w_l < w_u$). The inverse Fourier transform then yields the modulation filtered version of the auditory spectrogram. The bandpass modulation filter $H(w;[w_l,w_u])$ is defined as follows:

$$H(w;[w_l,w_u]) = (G^2) * exp(1-G^2)$$

where G is a montonically non-decreasing function given by $G = g(w;[w_l,w_u]) = mw$, with

$$m = \begin{cases} 1/w_l & \text{for } 0 <= w < w_l \\ 1/w & \text{for } w_l <= w <= w_u \\ 1/w_u & \text{for } w_u < w <= w_r \end{cases}$$

where w_r is the modulation frequency resolution and w_l, w_u are the lower and upper frequency cutoffs for a given bandpass modulation frequency range. With 10ms frame rate and 6 frequency channels per octave used in the auditory spectrogram computation, w_r is 50Hz for temporal modulations and 3 Cycles/Octave for spectral modulations. Note that for $w_l = 0$, the filter is lowpass.

In this work, we defined 4 different feature streams using two ranges of spectral and temporal modulations (shown in Table 1). The ranges are chosen carefully considering three important aspects; (i) each stream needs to carry *sufficient* information about the underlying signal (ii) there is *complimentary* information between different streams in terms of signal encoding (iii) constraining modulation bandpass cutoffs to ranges shown to be crucial for speech comprehension and highly robust to noise [10]. The aspects (i) and (ii) are crucial when combining information from the different feature streams, while (iii) is crucial to obtain high overall noise robustness performance. Figure 1 shows the four different feature streams for an example speech utterance.

Table 1: Range of spectral and temporal modulations captured by each of the 4 streams

Stream No.	Spectral modulations (Cycles/Octave)	Temporal modulations (Hz)
1	0 to 1 (slow)	0.5 to 12 (slow)
2	0.5 to 2 (fast)	0.5 to 12 (slow)
3	0 to 1 (slow)	10 to 22 (fast)
4	0.5 to 2 (fast)	10 to 22 (fast)

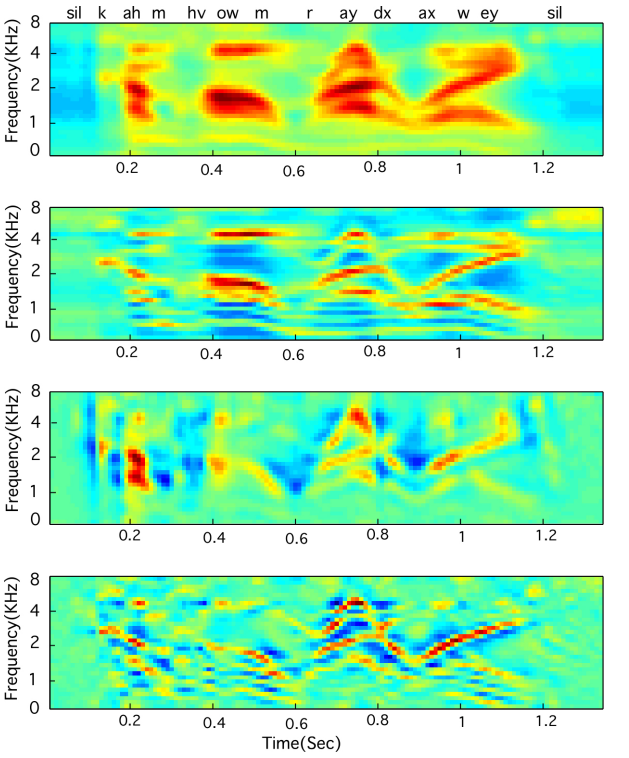

Figure 1: Illustration of the four different feature streams for the utterance "*come home right away*" taken from TIMIT speech database. The ordering of the streams 1-4 is from top to bottom. The top panel also shows the underlying phoneme label sequence.

3. Experiments and results

3.1. Recognition setup

Speaker independent phoneme recognition experiments are conducted on TIMIT database (excluding 'sa' dialect sentences), using the hybrid Hidden Markov Model / Multilayer perceptron (HMM/MLP) framework [11]. The training, cross-validation and test sets consist of 3400, 296 and 1344 utterances from 375, 87 and 168 speakers respectively. For the purpose of training and decoding, 61 hand-labeled symbols of the TIMIT training transcription are mapped to a standard set of 39 phonemes along with an additional garbage class [12].

MLP with a single hidden layer is trained to estimate the posterior probabilities of phonemes (conditioned on the input acoustic feature vector) by minimizing the cross entropy between the input feature vectors and the corresponding phoneme target classes [13]. These estimates are further refined by training a second MLP (in a hierarchical fashion) [14] which operates on a longer temporal context of 23 frames of posterior probabilities estimated by the first MLP. Both MLPs have a single hidden layer with sigmoid nonlinearity (1500 hidden nodes) and an output layer with softmax nonlinearity (40 output nodes). In the proposed multistream system, phoneme posteriors obtained from individual streams are integrated using a combination rule

Table 2: TIMIT ASR results in terms of phoneme recognition rate (in percentage) for different types of noise

Noise Type	SNR (in dB)	Features				
		MFCC39	MFCC+9FTC	MVA	ETSI	Multistream
Clean	∞	70.1	71.4	68.2	70.6	72.7
Factory1	20	48.5	48.2	55.7	61.5	66.4
	15	38.8	38.1	48.4	54.9	60.3
	10	27.2	28.3	39.4	45.1	51.3
	5	16.1	19.5	30.2	34.5	38.4
	Average	32.6	33.5	43.4	49	54.1
Babble	20	48.5	48.1	56.5	62.1	68.4
	15	36.6	37.3	49.5	55.6	62.9
	10	26.6	27.6	40.7	46.1	52.9
	5	18.4	19.5	29.7	34	36.9
	Average	32.5	33.1	44.1	49.4	55.3
Volvo	20	60.4	60.8	63.5	68.1	72.5
	15	55.6	55.7	62	66.7	72.4
	10	50.1	49.9	60.2	64.8	72.1
	5	41.8	42.9	58.1	61.7	71.1
	Average	51.9	52.3	60.9	65.3	72
F16	20	48.3	48.5	57.1	63.3	66.7
	15	37.3	37.8	50.8	57.9	61.1
	10	24.7	27	43.2	49.4	51.9
	5	14.3	18.2	34.6	38.5	40.3
	Average	31.1	32.9	46.4	52.3	55

based on the Dempster-Shafer (DS) theory of evidence [15] which has been shown to be related to human way of processing multiple feature streams. The final posterior probability estimates are converted to scaled likelihoods by dividing them with the corresponding prior probabilities (unigram language model) of phonemes. An HMM with 3 states, with equal self and transition probabilities associated with each state, is used for modeling each phoneme. The emission likelihood of each state is set to the scaled likelihood. Finally, the Viterbi algorithm is applied for decoding the phoneme sequence. Note that the hybrid HMM/MLP system achieves better phoneme recognition performance than the standard HMM/GMM systems [16].

3.2. Recognition results

The phoneme recognition performance for the proposed multistream system is compared against the performance obtained with a standard baseline features and two state-of-the-art noise robust feature schemes. The baseline features are MFCC39, obtained by taking the standard 13 Mel frequency cepstral coefficients along with their first and second order dynamic features (standard 39 dimensional MFCC features). A modified version of the baseline features, referred to as MFCC+9FTC, are obtained by taking a 9-frame temporal context on the standard 13 Mel frequency cepstral coefficients along with their first, second, and third order dynamic features (dimensionality is 9 x 13 x 4 = 468). Note that the modified version of the baseline features improves over the standard MFCC39 features in the hybrid HMM/MLP recognition framework. The first noise robust feature scheme compared against is Mean-Variance ARMA (MVA) processing of MFCC features [7]. The MVA processing is applied on the MFCC+9FTC features, and it combines the advantages of multiple noise robustness schemes: cepstral mean subtraction, variance normalization, and temporal filtering techniques like RASTA [17]. The second robust feature scheme compared against is the Advanced-ETSI distributed speech recognition front-end [8]. A 9-frame temporal context is taken on the ETSI features along with their first, second, and third order dynamic features, resulting in an input feature dimensionality of 468[1]. Both MVA and ETSI have been shown to provide excellent robustness for additive noise distortions, and form the state-of-the-art in noise robust feature schemes. For the multistream features, a 3-frame temporal context is taken on the base features along with their first, second, and third order dynamic features, resulting in an input feature dimensionality of 384 (3 x 32 x 4).

To evaluate the noise robustness aspect of the different feature representations, various noisy versions of the test set are created by adding four types of noise at Signal-to-Noise-Ratio (SNR) levels of 20dB, 15dB, 10dB and 5dB. The noise types chosen are, Factory floor noise (Factory1), Speech babble noise (Babble), Volvo car interior noise (Volvo), and F16 cockpit noise (F16), all taken from NOISEX-92 database [18], and added using the standard FaNT tool [19]. In all the experiments, the recognition models are trained only on the original clean training set and tested on the clean as well as noisy versions of test set (*mismatch* train and test conditions).

The phoneme recognition accuracy of various features is listed in Table 2. The proposed multistream features achieve performance comparable (or better) to that of MFCC, MVA, and ETSI features, under clean (matched) conditions. With additive noise conditions reflecting a variety of real acoustic scenarios, the multistream features perform substantially better than the baseline MFCC features, and significantly better than the

[1]For both ETSI and MFCC, the 9 frame context window and the 468 dimensional feature representations achieved best ASR performance

MVA and the ETSI features; an average relative improvement of 55.7%, 21.3%, and 9.5%, respectively. Note that the ETSI features have an additional advantage of using voice activity detectors (VAD) to identify noise-only frames and use the information to enhance the signal representation.

4. Discussion

In this work, we propose multistream bandpass modulation features for noise robust speech recognition. Multiple streams of features are constructed based on bandpass modulation filtering, with each stream covering an entire range of temporal and spectral modulations within carefully chosen modulation frequency bands. The rationale for multistream processing of speech features is highly grounded in the functional neuroanatomy of sound processing in the brain. Not only are neuron ensembles tuned to multiple scales of spectro-temporal modulations in the signal [20], there is also neurophysiological evidence of multiple streams of sound processing in the temporal lobe [1, 21]. The proposed scheme builds on this notion of parallel paths of processing to carefully devise four processing paths, divided along a slow vs. fast duality in both spectral and temporal modulations. Each stream provides a complimentary view of the modulation content in the speech signal, highlighting its fast vs. slow temporal dynamics or fast vs. slow spectral variations. This bisection of the modulation space is motivated by the duality of processing time constants in speech signals; as well as the contrast between subharmonic and broad spectral information. Perceptual studies have shown that concurrent streams of speech information along this dual slow vs. fast divide add up supra-linearly, leading to improved intelligibility relative to each stream by itself [22]. Such result is also observed here in our ASR experiments, with significant improvements over state-of-the art systems, both in clean and noise/mismatch conditions[2].

Note that in this work, we have not used any noise compensation techniques that involve voice/speech activity detectors to identify noise-dominated frames or subtract the noise component from the signal representations. The noise robustness improvements are purely due to the underlying feature representations, and are applicable to cases where noise compensation and/or signal enhancement techniques are not practical or feasible. However, we anticipate further improvements in the robustness by applying the proposed multistream approach on enhanced signal representations obtained from speech enhancement techniques [23]. It is also worth noting that the noise robustness obtained here from the multistream framework on a hybrid HMM/MLP system could be easily extended to other large scale speech recognition tasks in the TANDEM framework [24].

5. Acknowledgment

This research is partly supported by grants IIS-0846112 (NSF), FA9550-09-1-0234 (AFOSR), 1R01AG036424-01 (NIH) and N000141010278 (ONR).

6. References

[1] G. Hickock and D. Poeppel, "The cortical organization of speech processing," *Nature neurosc. reviews*, vol. 8, pp. 393–402, 2007.

[2] K. Wang and S. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Sp. Aud. Process.*, vol. 3, pp. 382–395, 1995.

[3] O'Connor KN, Yin P, Petkov CI, and Sutter ML, "Complex spectral interactions encoded by auditory cortical neurons: Relationship between bandwidth and pattern," *Front Syst Neurosci*, vol. 4, 2010.

[4] J.Woojay and B.H.Juang, "Speech analysis in a model of the central auditory system," *IEEE Trans. Sp. Aud. Process.*, vol. 15, pp. 1802–1817, 2007.

[5] S. Y. Zhao, Ravuri S., and Morgan N., "Multi-stream to many-stream: Using spectro-temporal features for asr," *In Proc. of INTERSPEECH*, pp. 2951–2954, 2009.

[6] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition," *In Proc. of INTERSPEECH*, pp. 318–321, 2010.

[7] Chia-Ping Chen and Jeff A. Bilmes, "Mva processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[8] ETSI, "Etsi es 202 050 v1.1.1 stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[9] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, pp. 824–839, 1992.

[10] T. Elliott and F. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput Biol*, vol. 5, pp. e1000302, 2009.

[11] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, Kluwer Pub., 1994.

[12] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust. Sp. Sig. Process.*, vol. 37, pp. 1641–1648, 1989.

[13] M.D. Richard and R.P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[14] J. Pinto, Sivaram G.S.V.S, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator," *IEEE Trans. Speech and Audio Process.*, vol. 19, pp. 225–241, 2011.

[15] Fabio Valente, "Multi-stream speech recognition based on dempster-shafer combination rule," *Speech Comm*, vol. 52(3), pp. 213–222, 2010.

[16] G.S.V.S. Sivaram, Nemala Sridhar Krishna, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *Signal Processing Letters, IEEE*, vol. 17, no. 11, pp. 957 –960, nov. 2010.

[17] H. Hermansky and Nelson Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 4, pp. 382–395, 1994.

[18] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," *Tech. Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K.*, 1992.

[19] H.G. Hirsch, "FaNT: Filtering and Noise Adding Tool," http://dnt.kr.hsnr.de/download.html.

[20] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[21] J. P. Rauschecker, "Cortical processing of complex sounds," *Curr. Opin. Neurobiol.*, vol. 8, pp. 516–521, 1998.

[22] M. Chait, S. Greenberg, T. Arai, J.Z. Simon, and D. Poeppel, "Two time scales in speech processing," in *Annual Meeting of the Cognitive Neuroscience Society, New York, NY*, 2005.

[23] P. C. Loizou, *Speech Enhancement: Theory and Practice*, (Boca Raton, FL), 2007.

[24] H. Hermansky, Daniel P.W. Ellis, and S. Sharma, "Tandem connection- ist feature extraction for conventional hmm systems," *In Proc. of ICASSP*, 2000.

[2] A detailed analysis of the contribition of each stream to the ASR performance and robustness shall be described in a future publication