A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition

Sridhar Krishna Nemala, Kailash Patil, and Mounya Elhilali, *Member, IEEE*

***Abstract*—There is strong neurophysiological evidence suggesting that processing of speech signals in the brain happens along parallel paths which encode complementary information in the signal. These parallel streams are organized around a duality of slow vs. fast: Coarse signal dynamics appear to be processed separately from rapidly changing modulations both in the spectral and temporal dimensions. We adapt such duality in a multistream framework for robust speaker-independent phoneme recognition. 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 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 and state-of-the-art feature schemes for phoneme recognition, particularly in presence of nonstationary noise, reverberation and channel distortions.**

***Index Terms*—Auditory cortex, automatic speech recognition (ASR), modulation, multistream, speech parameterization.**

I. Introduction

AUTOMATIC speech recognition (ASR) systems suffer a significant drop in performance when there is a mismatch between field test data and corpora used in system training. Mismatches are often caused by ambient background conditions or transmission channel distortions; which introduce both additive noise signals (at varying signal-to-noise levels) or additional linear or nonlinear distortions. Such distortions often have dramatic effects on the performance of ASR systems, even when the levels of mismatches are quite low. The noise robustness issue or ability to deal with mismatch train/test acoustic conditions is all the more important as the ASR technology is increasingly used for data input in mobile devices, where the acoustic input is expected to come from a diverse set of acoustic environments and channel conditions.

Given that humans are quite adept at communicating even at relatively high levels of noise [1], numerous noise robustness approaches in the literature focused on bringing biological intuition and knowledge into front-end feature extraction (e.g. Mel cepstral analysis, perceptual linear prediction, RASTA filtering) [2]. Additional techniques were also proposed to address mismatches due to stationary or slow-varying noise/channel; such as spectral subtraction (SS)[3], log-DFT mean normalization (LDMN), long-term log spectral subtraction (LTLSS), cepstral mean normalization (CMN) [4], and variance normalization [5]. Overall, state-of-the-art systems rely on a combination of auditory motivated front-end schemes augmented with feature normalization techniques [6]. Such schemes are often augmented with speech enhancement front-ends in order to tackle mismatch conditions [7], [8]. Nonetheless, current techniques remain quite limited in dealing with various classes of distortions, including nonstationary noise sources, reverberant noises and slowly-varying channel conditions.

To improve noise robustness, an alternative technique based on multistream combination was proposed [9], [10]. In this approach, multiple feature representations of the signal are processed in parallel before the information is combined at a later point. The motivation behind the multistream approach is that any external distortion does not affect the different feature streams in the same way, and by combining the information from multiple sources, the recognition performance can be improved. Traditionally, feature representations based on short-term analysis are combined with features that integrate information over long time windows [9]–[11]. Recently, feature schemes that take advantage of multiscale spectro-temporal modulations have been proposed; whereby conventional Gabor filters centered at a number of specific spectral or temporal modulation frequencies are used [12], [13]. These schemes have the drawback of dimensionality explosion of the feature space into several thousands of dimensions [14], [15], and typically use the rationale of feature division by segmenting the feature space into several tens of feature streams [12], [13], [16]–[18].

In this paper, we present a new framework for multistream feature processing. The proposed scheme puts much attention into the careful design of the multiple processing paths, taking into account a number of important points: **(i)** to integrate slow and fast dynamics of speech, both spectrally and temporally; **(ii)** to make each stream by itself noise-robust; **(iii)** to allow each

Manuscript received March 23, 2012; revised July 25, 2012; accepted August 31, 2012. Date of publication September 18, 2012; date of current version December 21, 2012. This work was supported by grants from the National Science Foundation (IIS CAREER 0846112), the Air Force Office for Scientific Review (FA9550-09-1-0234), the National Institutes of Health (1R01AG036424), and the Office of Naval Research (MURI N000141010278). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dong Yu.

The authors are with the Department of Electrical and Computer Engineering, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mounya@jhu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2219526

processing stream to cover an expanded range of spectral and temporal dynamics of speech. A key component that makes this approach feasible is the use of bandpass modulation filters; unlike the conventional Gabor filters that are *localized* around a specific spectral or temporal modulation frequency. As such, the proposed scheme avoids the drawbacks of previous attempts at multistream processing. We evaluate the benefits of this framework, on a variety of mismatch train and test conditions, in a speaker-independent phoneme recognition task. Preliminary results of this model were presented in [19].

The following section presents motivation and details of the multistream feature parameterization. The ASR system setup and the extensive set of recognition experiments involving a multitude of mismatch conditions are described in Section III. In Section IV, we contrast the recognition performance of the multistream system 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 [6] and Advanced-ETSI noise-robust speech recognition front-end [7]. An elaborate discussion on the proposed multistream parameterization, along with individual feature stream performances, is presented in Section V. We finally conclude with a brief summary and potential improvements towards achieving further robustness to noise distortions in Section VI.

II. Multistream Parameterization

The parameterization of speech sounds is achieved through a multistage auditory analysis that captures processing taking place at various stages along the auditory pathway from the periphery all the way to the primary auditory cortex (A1). We first describe the peripheral analysis used to obtain the 'auditory spectrogram' representation, followed by a detailed description of the cortical analysis for multistream parameterization of the speech input.

A. Peripheral Analysis

The acoustic input undergoes a series of transformations in the auditory periphery and is converted from a one-dimensional time waveform to a two-dimensional pattern of time-varying neuronal responses distributed along the tonotopic frequency axis. The two-dimensional representation referred to as the *auditory spectrogram* is obtained using an auditory-inspired model of cochlear and midbrain processing [20]. The acoustic input $s(t)$ is first processed through a pre-emphasis stage, implemented as a first-order highpass filter with pre-emphasis coefficient 0.97. An affine wavelet transform of the acoustic signal $s(t)$ models the spectral analysis observed at the cochlear stage. The cochlear frequency analysis is modeled by a bank of highly-asymmetric and overlapping constant-Q ($Q = 4$) bandpass filters ($h(t; f)$) with center frequencies that are uniformly distributed along a logarithmic frequency axis (x) ((1a), where $\otimes_t$ denotes convolution with respect to time). This stage employs 128 filters over a 5.3 octave range (24 filters/octave), and results in a spatiotemporal pattern of displacements $y_{coch}(t, f)$ along the basilar membrane. The following stage simulates the function of a lateral inhibitory network (LIN) which detects discontinuities in the responses across the tonotopic axis of the auditory nerve array, inducing a sharpening of the filterbank frequency selectivity as observed in the cochlear nucleus [20]. The LIN is approximated by a first derivative with respect to the tonotopic axis, followed by a half-wave rectifier (1b) and a short-term integrator (1c). The temporal integration window is implemented by the function $\mu(t; \tau) = e^{-t/\tau}u(t)$ with time constant $\tau = 10\ \mathrm{ms}$ mimicking the further loss of phase-locking observed in the midbrain. This stage effectively sharpens the bandwidths of the cochlear filters from $Q \simeq 4$ to 12. The final stage consists of a nonlinear cubic root compression of the spectrogram (1d), followed by downsampling the number of frequency channels by a factor 4, resulting in 32 frequency channels with a resolution of 6 channels/octave over 5.3 octaves.

$$y_{coch}(t, f) = s(t) \otimes_t h(t; f) \tag{1a}$$

$$y_{lin}(t, f) = max\left(\partial_f y_{coch}(t, f), 0\right) \tag{1b}$$

$$y_{mid}(t, f) = y_{lin}(t, f) \otimes_t \mu(t; \tau) \tag{1c}$$

$$y(t, f) = \left(y_{mid}(t, f)\right)^{\frac{1}{3}} \tag{1d}$$

B. Cortical Analysis

The spectrogram representation of a typical speech utterance is rich in temporal and frequency patterns, with fluctuations of energy across both time and frequency. These energy fluctuations, referred to as modulations, characterize several important cues and features associated with different sound percepts. Slow temporal modulations (<10 Hz) are commensurate with the syllable rate in speech, while intermediate and fast modulation rates (>10 Hz) capture segmental transitions like onsets and offsets. Similarly, slow/broad spectral modulations (<1 cycles/octave) capture primarily the overall spectral profile and formants, while fast/narrow modulation scales (>1 cycles/octave) reflect spectral details such as harmonics and subharmonic structure of the spectrum. Higher central auditory stages, especially the primary auditory cortex (A1), analyze the auditory spectrogram into more elaborate representations that highlight the spectro-temporal modulations present in the signal. Physiological data indicates that individual neurons in the central auditory pathway are tuned not only to frequencies but also selective to different ranges of spectral modulations (also referred to as scales) and temporal modulations (also referred to as rates) [21]. Furthermore, neural processing in the temporal lobe is functionally organized along parallel pathway, with dual streams attuned to either slow or fast dynamics of the sound signal [22], [23]. We mimic this cortical processing, and propose a multistream feature framework that integrates slow and fast dynamics of speech, both spectrally and temporally.

Each individual feature stream is obtained by filtering the auditory spectrogram $y(t, f)$ to capture different ranges of spectral and temporal modulations. The filtering process is performed in the Fourier domain on the modulation amplitudes using a set of bandpass spectral and temporal modulation filters. First, the Fourier transform of each spectral (or temporal) slice in the spectrogram is taken, then is multiplied by a bandpass modulation filter $H_S(w; [w_l, w_u])$ (or $H_R(w; [w_l, w_u])$) capturing modulation content within the specified range $[w_l, w_u]$

$(w_l < w_u)$. The bandpass modulation filters $H_S(w;[w_l,w_u])$ and $H_R(w;[w_l,w_u])$ are defined as follows:

$$H_S\left(w;[w_l,w_h]\right) = (\alpha w)^8 e^{\left[4-(2\alpha w)^2\right]}, \tag{2}$$

$$H_R\left(w;[w_l,w_h]\right) = (\alpha w)^2 e^{\left[1-(\alpha w)^2\right]},$$

$$\alpha = \begin{cases} \frac{1}{w_l}, & 0 \le w < w_l \\ \frac{1}{w}, & w_l \le w \le w_h \\ \frac{1}{w_h}, & w_h < w \le w_{max}, \end{cases} \tag{3}$$

where w_l, w_u are the lower and upper frequency cutoffs for a given bandpass modulation frequency range and w_{max} is the modulation frequency resolution. With 10 ms frame rate and 6 frequency channels per octave used in the auditory spectrogram computation, w_{max} is 50 Hz for temporal modulations and 3 cycles/octave for spectral modulations. Note that the modulation filters capture *a range of* modulations, and can be bandpass, lowpass (for $w_l = 0$) or highpass (depending on the value of w_h and the filter roll-off). Also note that the filter shape is carefully designed to have a long roll off on the high-frequency end.

The auditory spectrogram computation and the subsequent modulation filtering for multistream feature computation are done utterance by utterance (which vary in length between 1–5 seconds). The inverse Fourier transform then yields the bandpass modulation filtered version of the auditory spectrogram, and the resulting 32 dimensional spectral representation for every 10 ms time frame are taken as the *stream-specific* base features. The modulation filtering is performed in the real domain with no alterations to the phase information. In our implementation, spectral filtering is performed first, followed by temporal filtering.[1]

A summary of the parametrization procedure is as follows:

- Map the acoustic waveform $s(t)$ into a time-frequency representation $y(t,f)$
- For each stream i, perform a sequential filtering operation (along spectral slices then temporal slices of the spectrogram); where the Fourier transform of each slice is bandpass-filtered (in the real domain) using filters $H_S(w;[w_l,w_u])$ (or $H_R(w;[w_l,w_u])$). The net effect of this sequence of filtering operations is a new representation of the spectrogram viewed through the lens of the specific spectral and temporal bandpass filter (Fig. 2).

C. Stream Definitions

We parameterize speech with three feature streams that are carefully defined based on the spectro-temporal modulation profile of speech. Each feature stream encodes a range of spectral and temporal modulations capturing slow and/or fast dynamics of speech along each dimension. The stream index and the modulation ranges are shown in Table I. The streams are designed considering the following three important principles:

(I) *Information encoding*: Each feature stream by itself needs to carry *sufficient* information about the underlying speech signal. In the spectro-temporal modulation space, this requirement translates to each stream capturing a reasonable percentage of the total modulation energy. The correspondence between modulations in speech signals and the phonetic identity of the sound is based on behavioral studies that tie the fidelity of spectrotemporal modulations its accurate perception by human listeners, especially in presence of noise [27]–[30]. Fig. 1 shows the average Modulation Power Spectrum (MPS) computed from the training set of the TIMIT database [31]. The speech MPS is obtained by averaging the power of each speech utterance's 2-D Fourier transform of the auditory spectrogram. The regions covered by the feature streams 1–3 are shown. In our definition of feature streams, each stream captures on average ~60% of the total modulation energy and importantly all three feature streams encode approximately equal amounts of signal energy.

(II) *Complimentary information*: There is complimentary information between different streams in terms of signal encoding. This requirement indirectly translates to each stream capturing slightly different (even if overlapping) information about the underlying signal. In this work, the three feature streams are defined to capture different regions in the spectro-temporal modulation space. Note that aspects **(I)** and **(II)** are crucial when combining information from different feature streams. We show in the results section that a simple combination of *evidence* from the three feature streams results in a significant

TABLE I
RANGE OF SPECTRAL AND TEMPORAL MODULATIONS CAPTURED BY EACH OF THE THREE FEATURE STREAMS

Stream No.	Spectral Modulations (cycles/octave)	Temporal Modulations (Hz)
1	0 to 1.2	0.5 to 12
2	0.4 to 2.2	0.5 to 16
3	0 to 1.5	6 to 22

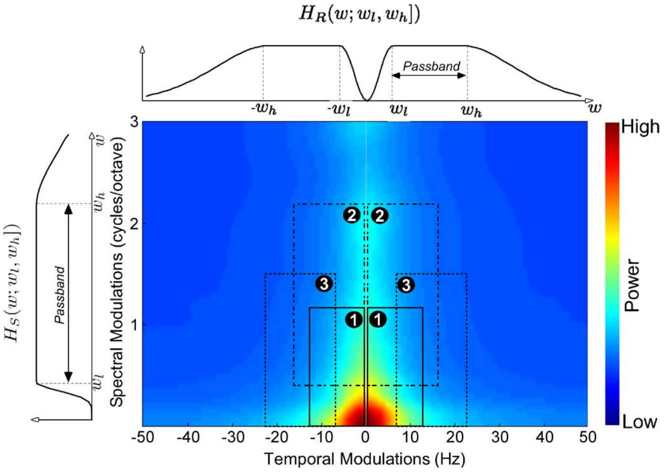

Fig. 1. The spectro-temporal average modulation power spectrum (MPS) of clean speech computed over the TIMIT corpus. The filter bandpass ranges for streams 1, 2 and 3 are overlayed over the speech MPS. Note that each rectangle does not reflect the entire range of the corresponding stream; since both spectral and temporal filters have long tails. As an example of filter shapes, the top panel shows the shape of the temporal modulation filter used for stream3 and leftmost panel shows the shape of the spectral modulation filter used for stream2.

[1] Switching the order of the filtering operation or performing 2-dimensional filtering over the spectro-temporal profile did not yield noticeable differences in the system performance. Perceptual data suggests that one can view the spectrotemporal modulation domain as a separable product of a purely temporal and purely spectral function [24]–[26].

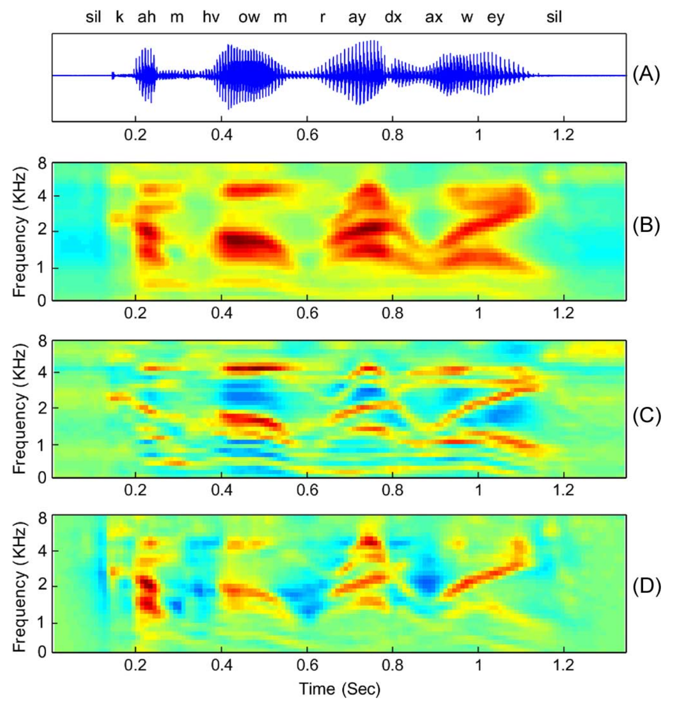

Fig. 2. Illustration of the three different feature streams for the utterance "*Come home right away*." taken from TIMIT speech database. The panel (A) shows the time domain waveform along with the underlying phoneme label sequence. Panels (B)–(D) show streams 1–3 in the same order.

improvement over any of the individual streams. Fig. 2 shows the three different feature streams for an example speech utterance.

(III) *Noise robustness*: Each feature stream by itself is noise-robust. We achieve this by constraining modulation bandpass cutoffs to ranges shown to be crucial for speech comprehension and highly robust to noise [30]. Each stream encodes high energy modulation components, which are inherently noise-robust, in the spectro-temporal modulation space.

III. Experiments

A. ASR System

An extensive set of speaker independent phoneme recognition experiments are conducted on TIMIT database using the hybrid Hidden Markov Model/Multilayer perceptron (HMM/MLP) framework [32]. The 'sa' dialect sentences are excluded from the experiments, as they might bias certain phoneme contexts and result in artificially high recognition scores [33]. The remaining training data of 3696 utterances is divided into sets of 3400 utterances from 375 speakers and 296 utterances from 87 speakers, and used for training and cross-validation respectively. The test data consists of an independent set of 1344 utterances from 168 speakers. 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 [33]. Note that the hybrid HMM/MLP framework overcomes some of the limitations of the standard HMM/GMM systems [32] and achieves better phoneme recognition performance [34] in addition to having advantages in dealing with noisy test data [35].

For each processing stream, a three layer MLP with a single hidden layer is discriminatively trained to estimate the posterior probabilities of phonemes (probability of phonemes conditioned on the input acoustic feature vector). The MLP weights are estimated using the training dataset by minimizing the cross entropy between the predicted posteriors and the corresponding phoneme target classes [36]. The cross-validation set is used to monitor learning progress and adjust the learning rate of the MLP using the standard back propagation algorithm. The posterior probability estimates are refined by training a second MLP, in a hierarchical fashion [37], which operates on a longer temporal context of 23 frames of posterior probabilities estimated by the first MLP. The contextual input (11 frames of left context, current frame, 11 frames of right context) allows the second MLP to use information about the correlations between successive feature vectors and improves the posterior estimates. 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 feature parameterization, phoneme posteriors for the three streams are estimated independently in parallel and finally combined using a simple product rule [11].

The final posterior probability estimates are converted to scaled likelihoods by dividing them with the corresponding prior probabilities of phonemes that are estimated from the relative frequencies of class labels in the training data. An HMM with 3 states, with equal self and transition probabilities associated with each state, is used for modeling each phoneme. The scaled likelihoods are used as the emission probabilities for the HMM during decoding. The standard Viterbi algorithm is applied for decoding the phoneme sequence. The recognition results on the test data are computed in terms of phoneme recognition rate by comparing (and aligning to compute insertions, deletions, and substitutions) the decoded phoneme sequence to the reference sequence of phonemes.

B. Mismatch Conditions

Two sets of recognition models are trained on clean TIMIT dataset; one on the original 16 kHz data and another on the downsampled 8 kHz data in order to match sampling in different noise corpora (as explained later). To evaluate the noise robustness aspect of the different feature representations, several noisy versions of the test set are created to simulate a number of real-world application scenarios. The following extensive set of noise types are evaluated in the experiments:

(i) Additive noise: Twenty versions of the test set (each 1344 utterances) are created by adding five different types of noise to the clean test data at Signal-to-Noise-Ratio (SNR) levels of $20 - 5$ dB (in steps of 5 dB) using the setup described in [38]. The noise types chosen are, Factory floor noise (Factory1), Speech babble noise (Babble), Volvo car interior noise (Volvo), F16 cockpit noise (F16), and Leopard military tank noise (Tank), all taken from NOISEX-92 database [39], and added using the standard FaNT tool [40].

(ii) Reverberant noise: Reverberation conditions are simulated by convolving the original clean test set with Gaussian white noise with exponentially decaying envelope. Five versions of the test set (each 1344 utterances,

TABLE II
TIMIT ASR Results in Terms of Phoneme Recognition Rate (PRR, in Percentage) on Clean Speech (16 kHz Data and 8 kHz Downsampled Data), Speech Corrupted With Additive Noise (Average Performance for Five Noise Types at $20-5$ dB SNRs), Reverberant Speech (Average Performance for 7 Impulse Responses With RT_{60} Ranging From 100–500 ms), and Telephone Channel Speech (Average Performance for Nine HTIMIT Channel Conditions). Multistream Feature Parameterization is Compared Against the Standard and State-of-the-Art Feature Schemes

Noise Type	Speech Parameterizations			
	MFCC	MFCC+MVA	ETSI	Multistream
Clean (16kHz)	71.4	68.2	70.6	73.1
Clean (8kHz)	70.5	67.5	69.3	71.3
Additive	39.2	49.4	54.8	60.8
Reverberant	31.6	34.6	33.8	40.9
Telephone Channel	36.4	54.5	52.7	56.4

labeled SR100 to SR500) are created with simulated reverberation time (RT_{60}) ranging from 100–500 ms (in steps of 100 ms). Two additional versions of test set (labeled RR100 and RR500) are created by convolving the test set with real room responses taken from [41] with reverberation time ~100 ms and ~500 ms.

(iii) Channel mismatch: Speech data from nine different telephone channels in the handset TIMIT (HTIMIT) [42] are used. The nine test sets contain 842 utterances each (with intersection to the original clean TIMIT test set, subset of the original 1344 utterances). The nine versions of the test are labeled with the original HTIMIT telephone channel labels (CB1, CB2, CB3, CB4, EL1, EL2, EL3, EL4, and PT1).

IV. Results

In all the recognition experiments, the models are trained only on the original clean training set and tested on the clean as well as noisy versions of test set representing *mismatch* train and test cases. For the evaluation of additive and reverberant test noise conditions, recognition system trained on 16 kHz is used. For evaluating mismatch conditions involving telephone channel, recognition setup trained on 8 kHz data is used.

The phoneme recognition performance for the proposed multistream feature framework is compared against the performance obtained with a standard baseline features and two state-of-the-art noise robust feature schemes. The baseline features evaluated are standard Mel-Frequency Cepstral Coefficients (MFCC) [43]. MFCC features are obtained by stacking a set of 9 frames of standard 13 MFCCs along with their first, second, and third order temporal derivatives (dimensionality is $9 \times 13 \times 4 = 468$). Note that this modified version of the baseline features improves over the standard 39 dimensional MFCC features in the hybrid HMM/MLP recognition framework [19]. For the multistream parameterization, 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 $(3 \times 32 \times 4)$ for each stream.

The first noise robust feature scheme compared against is Mean-Variance ARMA (MVA) processing applied on MFCC features [6]. This system combines the advantages of multiple noise robustness schemes: cepstral mean subtraction, variance normalization, and temporal filtering techniques like RASTA [44]. The second robust feature scheme compared against is the Advanced-ETSI distributed speech recognition front-end [7]. 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. For both ETSI and MFCC, the 9 frame context window and the 468 dimensional feature representations achieved the best ASR performance. Both MFCC+MVA and ETSI have been shown to provide excellent robustness for a variety of noise distortions, and form the state-of-the-art in noise robust feature schemes.[2]

An overview of the recognition performance for the different speech parameterizations is shown in Table II. A detailed analysis of the performance of the different feature schemes on speech corrupted with additive noise, reverberant speech, and telephone channel are shown in Tables III, IV, and V respectively.

Note that the baseline MFCC features, though perform better than MFCC+MVA and ETSI on clean speech, are the most affected on all noise conditions. The multistream parameterization performs significantly better than all the other feature schemes on clean as well as on all the mismatch train and test conditions. ETSI features perform significantly better than MFCC+MVA on additive noise conditions, while MFCC+MVA features perform better than ETSI on reverberant and telephone speech.

On speech corrupted with additive noise reflecting a variety of real acoustic background conditions, the multistream parameterization performs significantly better than the MFCC+MVA and the ETSI features; an average relative improvement of 23.1%, and 10.9%, respectively (averaged over the five noise types and 4 SNR conditions). The multistream parameterization gives an average relative improvement of 18.2% and 21% over MFCC+MVA and ETSI features on reverberant speech (averaged over the seven reverberation conditions), and an average relative improvement of 3.5% and 7% over MFCC+MVA and ETSI features on the telephone speech (averaged over the nine HTIMIT channel conditions). Note that, of the feature schemes evaluated, ETSI features have an additional advantage of using voice/speech activity detectors (VAD) to identify noise-only frames and use the information to enhance the signal representation.

V. Analysis

A. Individual Stream Performance

The average recognition performance of the three feature streams, as a percentage of the multistream combination performance, for different noise conditions is shown in Fig. 3. The complete set of stream-specific recognition results for all the noise conditions is given in the Appendix. It can be readily seen that all the streams contribute at least more than 80%, and ~90% on an average, to the combination performance.

[2] The ETSI and MFCC+MVA features, on a comparable phoneme recognition task, have shown to be comparable or better than several other noise robust feature schemes namely modulation spectrum based features, relative spectral (RASTA) filtering and multi-resolution RASTA (MRASTA), gammatone frequency cepstral coefficients (GFCC), log-DFT mean normalization (LDMN), and long-term log spectral subtraction (LTLSS) [45].

TABLE III
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON SPEECH CORRUPTED WITH A VARIETY OF ADDITIVE NOISE TYPES. MULTISTREAM FEATURE PARAMETERIZATION IS COMPARED AGAINST CONVENTIONAL AND STATE-OF-THE-ART NOISE-ROBUST FEATURE SCHEMES

Noise Type	SNR (in dB)	Speech Parameterizations			
		MFCC	MFCC+MVA	ETSI	Multistream
Babble	20	48.1	56.5	62.1	68.0
	15	37.3	49.5	55.6	62.7
	10	27.6	40.7	46.1	53.0
	5	19.5	29.7	34.0	38.4
	Average	33.1	44.1	49.5	55.5
F16	20	48.5	57.1	63.3	66.5
	15	37.8	50.8	57.9	60.6
	10	27.0	43.2	49.4	51.2
	5	18.2	34.6	38.5	40.1
	Average	32.9	46.4	52.3	54.6
Volvo	20	60.8	63.5	68.1	73.0
	15	55.7	62.0	66.7	72.8
	10	49.9	60.2	64.8	72.4
	5	42.9	58.1	61.7	71.6
	Average	52.3	61.0	65.3	72.5
Factory1	20	48.2	55.7	61.5	66.1
	15	38.1	48.4	54.9	59.5
	10	28.3	39.4	45.1	50.3
	5	19.6	30.2	34.5	38.7
	Average	33.6	43.4	49.0	53.6
Tank	20	54.3	57.8	62.8	71.6
	15	47.9	54.5	59.7	70.1
	10	40.4	50.7	56.9	66.8
	5	32.9	46.4	51.7	62.1
	Average	43.9	52.3	57.8	67.7

TABLE IV
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON REVERBERANT SPEECH. MULTISTREAM FEATURE PARAMETERIZATION IS COMPARED AGAINST THE STANDARD AND STATE-OF-THE-ART NOISE-ROBUST FEATURE SCHEMES

Noise Type	Speech Parameterizations			
	MFCC	MFCC+MVA	ETSI	Multistream
SR100	47.7	50.1	48.5	56.8
RR100	36.6	48.4	44.1	57.3
SR200	35.5	37.3	36.4	43.4
SR300	29.6	30.5	30.5	36.4
SR400	26.3	27.1	27.2	32.8
SR500	24.5	24.6	25.1	30.1
RR500	21.1	24.0	24.5	29.4
Average	31.6	34.6	33.8	40.9

TABLE V
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON DIFFERENT TELEPHONE CHANNEL SPEECH (HTIMIT). MULTISTREAM FEATURE PARAMETERIZATION IS COMPARED AGAINST THE STANDARD AND STATE-OF-THE-ART NOISE-ROBUST FEATURE SCHEMES

Noise Type	Speech Parameterizations			
	MFCC	MFCC+MVA	ETSI	Multistream
CB1	48.9	58.9	59.5	62.7
CB2	42.7	61.6	61.7	66.1
CB3	27.1	43.3	44.4	38.4
CB4	32.4	48.5	45.5	47.0
EL1	49.3	62.0	60.9	65.2
EL2	31.3	55.9	52.8	57.2
EL3	37.1	52.9	50.8	57.4
EL4	30.7	56.7	47.6	58.0
PT1	27.8	50.7	51.4	55.8
Average	36.4	54.5	52.7	56.4

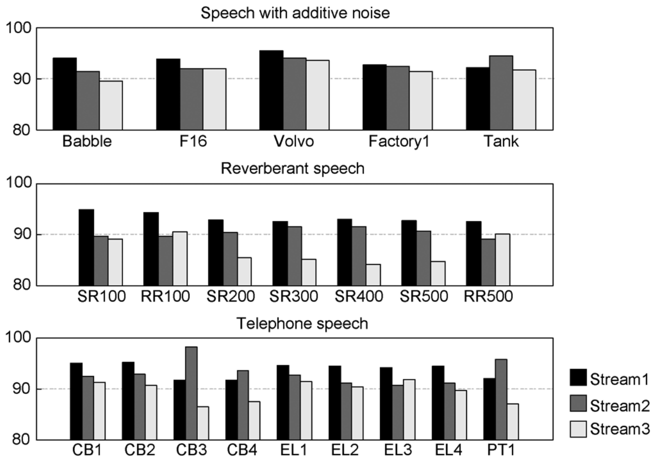

Fig. 3. Performance of the three feature streams, as a percentage of the combination performance, on speech with additive noise (top panel, each noise type performance is averaged over the SNRs 20, 15, 10, and 5 dB), reverberant speech (middle panel), telephone speech (bottom panel). Within each group, the bars from left to right are stream1, stream2, and stream3 respectively. The performance of the three streams in clean is 70.4%, 68.5% and 68.5% for streams 1, 2 and 3 respectively.)

This is consistent with the approximately equal information encoding capacity (w.r.t. the speech signal representation in the spectro-temporal modulation space) of the feature streams. Stream 1 which encodes slow spectral and temporal modulations gives the best performance, amongst the three feature streams, on majority of the noise conditions (17 out of 21). The superior performance of stream 1 is not surprising and can be attributed to the importance of the slow modulations for speech comprehension [30]. In fact, stream 1 alone performs better than the MFCC+MVA and ETSI noise-robust feature schemes on majority of the noise conditions (15 out of 21).

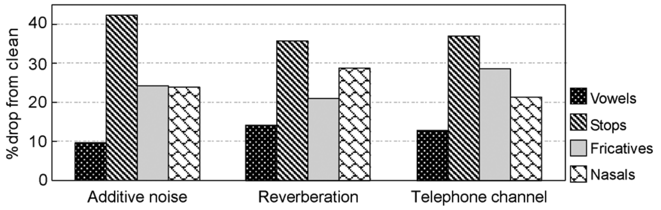

Fig. 4. Percentage drop (from clean) in recognition performance of broad phoneme classes due to additive noise, reverberation, and telephone channel; calculated from the average broad class phoneme recognition performace on the three noise types. Within each group, the bars from left to right are vowels, stops, fricatives, and nasals, respectively.

Though extended to include faster temporal modulations, streams 2 and 3 do still cover crucial parts of the lower modulation range due to the carefully chosen roll-off definitions of the modulation filters. Both streams give slightly lower but comparable noise-robustness performances, since the faster modulations encoded in the streams are still constrained to the ranges that are crucial to speech comprehension. Note that it is not entirely straight-forward to relate the ranges of modulations encoded in each feature stream to noise robustness (on average or even for certain noise conditions). While slow modulations may be advantageous in terms of robustness, they may still result in lower recognition performance since the encoding of only the slow dynamics may compromise phoneme discriminability. Similarly, including fast modulations (upto a certain extent) may result in a higher recognition performance because of the superior phoneme discriminability even though the robustness aspect may be compromised to some extent. The issue of phoneme discriminability is addressed next.

B. Broad Phoneme Class Recognition

We first highlight the observed effects of different noise types on broad class phoneme recognition. A subset of 39 phonemes are grouped into four broad phoneme classes: vowels, stops, fricatives, and nasals. To highlight the effect of different noise types, the percentage drop (from clean) in recognition performance of each of the phoneme classes is shown in Fig. 4. All the noise conditions significantly affect stops, while the vowels are the least affected. On average, fricatives and nasals are affected to the same degree; with nasals being more disrupted by reverberant speech and vice-versa with telephone speech. With stops accounting for more than 35% of the error relative to the other broad classes, increasing noise robustness for stops in particular would certainly benefit the overall robustness of any feature scheme. Such observation remains an open research direction.

We next highlight an interesting phenomenon in which parallel information across streams helps complement the contribution of each processing path in the entire scheme. Fig. 5 provides an example of such complementarity. The figure shows average broad class phoneme recognition performance of the three feature streams on clean speech and speech corrupted with additive noise. Notice that for stop consonants, on clean speech, the performance of all the feature streams is comparable with stream 1 having a slight advantage. However, on speech corrupted with additive noise, stream 2 gives a significantly higher performance as compared to stream 1 for the stop consonants.

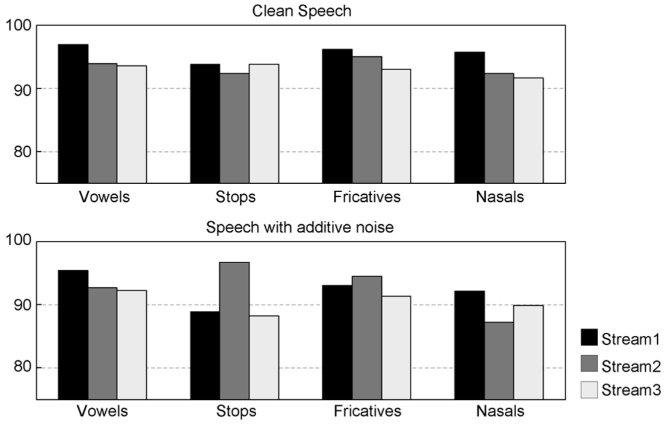

Fig. 5. The broad class phoneme recognition accuracies of the three feature streams, as a percentage of the combination performance, on clean speech (top panel) and speech with additive noise (bottom panel, averaged over all five noise types and four SNRs). Within each group, the bars from left to right are stream1, stream2, and stream3, respectively.

Psychophysical evidence suggests that one of the advantages that normal subjects have over hearing-impaired listeners is improved local target-to-masker ratios; especially in presence of non-stationary backgrounds [46]. The notion of listening in the spectral and temporal "dips" of the noisy signal is made possible for normal hearing listeners because of superior spectral selectivity and increased temporal resolution. This concept relates to the specific example of the improved recognition performance of stream 2, which encodes higher spectral and temporal resolutions, for stop consonants in the non-stationary additive noise conditions. However note that, due to the complex interaction of different noise types with speech signal space and their interaction together with back-end statistical models, it is not trivial in general to relate the performance of different feature schemes w.r.t. noise conditions.

C. Combination of Evidence

Perceptual studies have shown that concurrent streams of speech information along the proposed slow vs. fast divide add up supra-linearly, leading to improved intelligibility relative to each stream by itself [47]. In this work, we combined the three feature streams at the phoneme posterior level using a simple product rule [11]; the motivation being all the feature streams contribute to the signal encoding, due to the overlap regions in the ranges of modulations captured by the proposed bandpass modulation filters. The multistream combination results in an average (over all the noise conditions) relative improvement of 9.6% over the average individual stream recognition performance; highlighting the overlapping yet complimentary information captured by the three feature streams. The combination improvement is also significant considering that the noise robustness performance of each stream by itself is good. We also evaluated inverse entropy weighting and Demster-Shafer combination rules [11], with both giving comparable results to the product rule. Note that further gains in performance can be achieved by employing the "full combination (FC)" strategy [14], in which all possible combinations of feature streams are used by defining a set of exhaustive and mutually exclusive

TABLE VI
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) FOR TWO SUB-DIVISIONS OF STREAM1

Stream Definition	Noise Condition		
	Babble, 10dB	SR300	EL2
Stream1	49.2	33.7	54.0
Stream1_1	45.4	30.2	50.3
Stream1_2	42.8	29.1	48.6
Combination	47.3	31.8	51.9

experts. The combination strategies that make use of the noise characteristic estimates [48] can also significantly improve the noise robustness performance.

D. Are More Feature Streams Better?

We empirically show how a further sub-division of spectro-temporal modulation space into many feature streams does not necessarily result in an improved performance. We show this with an example by sub-dividing the modulation ranges captured by stream1 into two; with first sub-division labeled as Stream1_1 encoding 0–0.7 cycles/octave spectral and 0.5–8 Hz temporal modulations, and second sub-division labeled as Stream1_2 encoding 0.3–1.2 cycles/octave spectral and 4–12 Hz temporal modulations. The recognition performance of stream1, and its two sub-divisions and their combination, for three example noise conditions is given in Table VI. Notice that the performance of the two sub-stream combination is still inferior to the original stream. This might be due to the fact that each sub-stream by itself is not able to encode 'sufficient' information about the underlying speech signal, which is one of the key requirements in multistream feature design. The choice of these design principles is chiefly driven by the specific shape of our spectral and temporal filters; in terms of overlap between passbands, energy passed through stopband tails as well as roll-off factors. These parameters interact closely with our fusion rule across the three streams which does require them to provide *sufficient encoding power*. That being said, we have not exhaustively spanned all possibilities with filter shape, overlap and number of streams. It is also worth noting that improvements with further sub-division of the modulation space might be possible for certain artificial background conditions like localized ripple noise. The advances in combination techniques would also play a key role in multistream parameterizations.

VI. Summary and Conclusion

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. In this paper, we propose a novel multistream feature framework that integrates these slow and fast dynamics of speech. In the proposed scheme, three feature streams are carefully designed based on three design principles; (**i**) each feature stream by itself needs to carry *sufficient* information about the underlying speech signal (**ii**) there is complimentary information between different streams in terms of signal encoding (**iii**) each feature stream by itself is noise-robust.

We evaluate the benefits of the proposed framework on an extensive set of mismatch train and test conditions, and showed significant advantages as compared to state-of-the-art Mean-Variance normalization and ARMA filtering (MVA) and Advanced ETSI front-end noise robust feature schemes; an average relative improvement of the order of 15%. We further show that even a single stream (stream1) outperforms the state-of-the-art robust features on majority of the noise conditions (15 out of 21). However, if only a single feature stream was to be used, we would suggest a bandpass modulation filtering with ranges 0–2 cycles/octave for spectral modulations and 0.5–12 Hz for temporal modulations [49].

It is worth noting that the noise robustness improvements are obtained without any explicit feature normalization and/or signal enhancement techniques. Further improvements in robustness could be achieved by applying the multistream scheme on enhanced signal representations obtained from speech enhancement techniques [8] and/or applying various normalization techniques to minimize the effect of interference/noise in feature domain. The parameterization could also be *combined* with existing feature schemes to take advantage of any complimentary information between the representations. Such benefits can be readily observed when combining the multistream model proposed here with the MFCC+MVA or ETSI schemes (see Appendix, Table X). Fusing posterior estimates from these different schemes does in fact show evidence for improved robustness relative to each system by itself, and hence preliminary evidence of some degree of complimentarily between the different features. Alternative, more targeted, fusion methodologies could undoubtedly provide additional improvements.

A key component that makes the proposed multistream framework feasible and cover the entire spectro-temporal modulation space ('significant' energy region) with just three feature streams is the use of bandpass modulation filters which can span an entire range of either slow or fast spectral and temporal modulations. This also ensures no dimensionality expansion in the feature extraction stage, unlike the previous approaches encoding multiscale spectro-temporal modulations (that typically compute several thousands of feature dimensions), thereby making the proposed multistream parameterization also highly computationally efficient. The computational aspect is especially relevant in the context of mobile devices where the ASR technology is increasingly being used for data input. Note that even though the recognition in some cases is done on *cloud*, the feature extraction is still best done on the mobile device to avoid distortions introduced by channel used to transmit the signal (acoustic input).

The results presented in this work from the multistream parameterization on a hybrid HMM/MLP system could be extended to large scale speech recognition tasks in the TANDEM framework [50]. The parameterization could also be used to improve noise robustness in other automatic speech processing tasks. For example, speaker and language recognition, where it is common to use multiple feature representations and/or systems (back-end), can directly benefit this work; we have preliminary results that show great promise for speaker recognition.

APPENDIX

TABLE VII

Noise Type	SNR (in dB)	Stream Number			
		Stream1	Stream2	Stream3	Combination
Clean(16kHz)	-	70.4	68.5	68.5	73.1
Babble	20	64.5	63.6	63.1	68.0
	15	58.4	57.7	57.0	62.7
	10	49.2	47.9	46.2	53.0
	5	36.7	33.7	32.6	38.4
	Average	52.2	50.7	49.7	55.5
F16	20	63.3	61.8	62.1	66.5
	15	57.1	56.0	55.9	60.6
	10	47.8	47.2	46.7	51.2
	5	36.7	36.0	36.2	40.1
	Average	51.2	50.2	50.2	54.6
Volvo	20	70.1	68.5	68.4	73.0
	15	69.9	68.3	68.3	72.8
	10	69.1	68.1	67.8	72.4
	5	67.9	67.4	66.6	71.6
	Average	69.2	68.1	67.8	72.5
Factory1	20	62.3	61.8	61.2	66.1
	15	55.6	55.6	54.7	59.5
	10	46.0	46.3	45.4	50.3
	5	35.1	34.4	34.6	38.7
	Average	49.7	49.5	49.0	53.6
Tank	20	67.5	67.2	66.8	71.6
	15	65.3	66.0	64.7	70.1
	10	61.2	63.5	61.1	66.8
	5	55.6	59.1	55.9	62.1
	Average	62.4	63.9	62.1	67.7

TABLE VIII
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON REVERBERANT SPEECH FOR THE DIFFERENT PROCESSING STREAMS

Noise Type	Stream Number			
	Stream1	Stream2	Stream3	Combination
Clean(16kHz)	70.4	68.5	68.5	73.1
SR100	53.9	50.9	50.6	56.8
RR100	54.0	51.4	51.9	57.3
SR200	40.3	39.2	37.1	43.4
SR300	33.7	33.3	31.0	36.4
SR400	30.5	30.0	27.6	32.8
SR500	27.9	27.3	25.5	30.1
RR500	27.2	26.2	26.5	29.4
Average	38.2	36.9	35.7	40.9

TABLE IX
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON DIFFERENT TELEPHONE CHANNEL SPEECH (HTIMIT) FOR THE DIFFERENT PROCESSING STREAMS

Noise Type	Stream Number			
	Stream1	Stream2	Stream3	Combination
Clean(8kHz)	68.1	66.6	66.0	71.3
CB1	59.6	57.9	57.2	62.7
CB2	62.9	61.4	59.9	66.1
CB3	35.2	37.7	33.2	38.4
CB4	43.1	44.0	41.1	47.0
EL1	61.7	60.4	59.6	65.2
EL2	54.0	52.1	51.7	57.2
EL3	54.0	52.0	52.7	57.4
EL4	54.8	52.8	52.0	58.0
PT1	51.4	53.4	48.6	55.8
Average	52.9	52.4	50.7	56.4

TABLE X
TIMIT ASR RESULTS IN TERMS OF PHONEME RECOGNITION RATE (PRR, IN PERCENTAGE) ON CLEAN SPEECH (16 kHz DATA AND 8 kHz DOWNSAMPLED DATA), SPEECH CORRUPTED WITH ADDITIVE NOISE (AVERAGE PERFORMANCE FOR FIVE NOISE TYPES AT $20-5$ dB SNRS), REVERBERANT SPEECH (AVERAGE PERFORMANCE FOR 7 IMPULSE RESPONSES WITH RT_{60} RANGING FROM 100–500 ms), AND TELEPHONE CHANNEL SPEECH (AVERAGE PERFORMANCE FOR NINE HTIMIT CHANNEL CONDITIONS). MULTISTREAM FEATURE PARAMETERIZATION AND ITS COMBINATION WITH STATE-OF-THE-ART FEATURE SCHEMES ARE COMPARED HERE. COMBINATION OF EVIDENCE FROM DIFFERENT FEATURE PARAMETERIZATIONS ARE DONE USING THE PRODUCT RULE [11]

Noise Type	Speech Parameterizations		
	Multistream	Multistream + MFCC+MVA	Multistream + ETSI
Clean (16kHz)	73.1	74.6	74.6
Clean (8kHz)	71.3	73.1	73.2
Additive	60.8	61.8	62.0
Reverberant	40.9	42.2	41.1
Telephone Channel	56.4	59.6	53.1

REFERENCES

[1] S. Greenberg, A. Popper, and W. Ainsworth*, Speech Processing in the Auditory System*. Berlin, Germany: Springer, 2004.

[2] H. Hermansky, "Should recognizers have ears?," *Speech Commun.*, vol. 25, pp. 3–27, 1998.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic, Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[5] G. Cook, D. Kershaw, J. Christie, C. Seymour, and S. Waterhouse, "Transcription of broadcast television and radio news: The 1996 abbot system," *Proc. Int. Acoustics Speech Signal Process.*, pp. 723–726, 1997.

[6] C. Chen and J. Bilmes, "Mva processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[7] *ETSI ES 202 050 v1.1.1 STQ; Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 202 050 v1.1.1 STQ, ETSI, 2002.

[8] P. Loizou*, Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC, 2007.

[9] S. Sharma, "Multi-stream approach to robust speech recognition," Ph.D. dissertation, Oregon Graduate Inst. of Sci. Technol., Portland, OR, 1999.

[10] A. Hagen, "Robust speech recognition based on multi-stream processing," Ph.D. dissertation, Lab. Intell. Artif. Perceptive, cole Polytechnique Fdrale, Lausanne, Switzerland, 2001.

[11] F. Valente, "Multi-stream speech recognition based on Dempster-Shafer combination rule," *Speech Commun.*, vol. 52, no. 3, pp. 213–222, 2010.

[12] S. Y. Zhao, R. S. , and M. N. , "Multi-stream to many-stream: Using spectro-temporal features for ASR," in *Proc. INTERSPEECH*, 2009, pp. 2951–2954.

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

[14] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Commun.*, vol. 34, pp. 25–40, 2001.

[15] J. Woojay and B. Juang, "Speech analysis in a model of the central auditory system," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 6, pp. 1802–1817, Aug. 2007.

[16] M. Kleinschmidt, "Spectro-temporal gabor features as a front end for automatic speech recognition," in *Forum Acusticum*, 2002.

[17] B. Meyer, S. Ravuri, M. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," *Proc. INTERSPEECH*, vol. 1, pp. 1269–1272, 2011.

[18] H. Lei, B. Meyer, and N. Mirghafori, "Spectro-temporal gabor features for speaker recognition," in *Proc. IEEE Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4241–4244.

[19] S. Nemala, K. Patil, and M. Elhilali, "Multistream bandpass modulation features for robust speech recognition," in *Proc. ISCA*, 2011, pp. 1277–1280.

[20] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992.

[21] T. Chi, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.

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

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

[24] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *J Acoust Soc Amer.*, vol. 66, no. 5, pp. 1364–1380, Nov. 1979.

[25] D. Green*, Auditory Frequency Selectivity*. Cambridge, MA: Plenum, 1986, ch. Frequency and the detection of spectral shape change, pp. 351–359.

[26] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. A. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, pp. 2719–2732, 1999.

[27] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1053–1064, 1994.

[28] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, pp. 331–348, 2003.

[29] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. National Acad. Sci., USA*, vol. 102, no. 7, pp. 2293–2298, Feb. 2005.

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

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren*, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium, 1993, p. LDC93S1.

[32] H. Bourlard and N. Morgan*, Connectionist Speech Recognition: A Hybrid Approach*. Dordrecht, The Netherlands: Kluwer, 1994, p. 348.

[33] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[34] S. Garimella, S. Nemala, N. Mesgarani, and H. Hermansky, "Data-driven and feedback based spectro-temporal features for speech recognition," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 957–960, Nov. 2010.

[35] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 14–22, Jan. 2005.

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

[37] J. Pinto, S. Garimella, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analyzing MLP. Based Hierarchical Phoneme posterior probability estimator," *IEEE Trans. Speech and Audio Process.*, vol. 19, pp. 225–241, 2011.

[38] D. Gelbart, "Ensemble feature selection for multi-stream automatic speech recognition," Ph.D. dissertation, UC Berkeley, Berkeley, CA, 2008.

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

[40] H. Hirsch, FaNT: Filtering and Noise Adding Tool. [Online]. Available: http://dnt.kr.hsnr.de/download.html (date last viewed 11/25/2011), 2005

[41] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding (ASRU)*, 2001, pp. 103–106.

[42] D. Reynolds*, HTIMIT*. Philadelphia, PA: Linguistic Data Consortium, 1998, p. LDC98S67.

[43] *Readings in Speech Recognition*, A. Waibel and K. Lee, Eds. Burlington, MA: Morgan Kaufmann, 1990, p. 680.

[44] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 382–395, Oct. 1994.

[45] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *J. Acoust. Soc. Amer.*, vol. 128, pp. 3769–3780, 2010.

[46] M. K. Qin and A. J. Oxenham, "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Amer.*, vol. 114, no. 1, pp. 446–454, Jul. 2003.

[47] M. Chait, S. Greenberg, T. Arai, J. Simon, and D. Poeppel, "Two time scales in speech processing," in *Proc. Annu. Meeting Cognitive Neurosci. Soc.*, New York, 2005.

[48] N. Mesgarani, S. Thomas, and H. Hermansky, "Toward optimizing stream fusion in multistream recognition of speech," *J. Acoust. Soc. Amer.*, vol. 130, pp. 14–18, 2011.

[49] M. Carlin, K. Patil, S. Nemala, and M. Elhilali, "Robust phoneme recognition based on biomimetic speech contours," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH*, 2012.

[50] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1635–1638.

Sridhar Krishna Nemala received the Ph.D. degree in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, MD, in 2012. From 2004 to 2006, he was a research engineer with Hewlett-Packard Labs, where he worked on text-to-speech synthesis and handwriting/gesture recognition (in Language Technology and Applications department). He was recognized for outstanding contribution to HP Labs India achievements for the year 2004. Prior to that, he received M.S. (by Research) in computer science and engineering from Indian Institute of Technology—Madras. His research interests include machine learning and all aspects of audio, speech, and image processing.

Kailash Patil received his M.S.E. degree in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, in 2011. He is currently a doctoral student at the same department. He received his B. Tech degree in Electronics and Communication Engineering from Indian Institute of Technology, Guwahati, India, in 2008. His research interests include auditory scene analysis, speech processing and machine learning.

Mounya Elhilali (M’00) received her Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park, in 2004. She is now an assistant professor at the Department of Electrical and Computer Engineering at the Johns Hopkins University. She is affiliated with the Center for Speech and Language Processing and directs the Laboratory for computational Audio Perception. Her research examines the neural and computational bases of sound and speech perception in complex acoustic environments; with a focus on robust representation of sensory information in noisy soundscapes, problems of auditory scene analysis and segregation as well as the role of top-down adaptive processes of attention, expectations and contextual information in guiding sound perception. Dr. Elhilali is the recipient of the National Science Foundation CAREER award and the Office of Naval Research Young Investigator award.