Auditory Model-Based Design and Optimization of Feature Vectors for Automatic Speech Recognition

Saikat Chatterjee, *Member, IEEE*, and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—Using spectral and spectro-temporal auditory models along with perturbation-based analysis, we develop a new framework to optimize a feature vector such that it emulates the behavior of the human auditory system. The optimization is carried out in an offline manner based on the conjecture that the local geometries of the feature vector domain and the perceptual auditory domain should be similar. Using this principle along with a static spectral auditory model, we modify and optimize the static spectral mel frequency cepstral coefficients (MFCCs) without considering any feedback from the speech recognition system. We then extend the work to include spectro-temporal auditory properties into designing a new dynamic spectro-temporal feature vector. Using a spectro-temporal auditory model, we design and optimize the dynamic feature vector to incorporate the behavior of human auditory response across time and frequency. We show that a significant improvement in automatic speech recognition (ASR) performance is obtained for any environmental condition, clean as well as noisy.**

***Index Terms*—Auditory models, mel frequency cepstral coefficient (MFCC), speech recognition.**

I. Introduction

CURRENT automatic speech recognition (ASR) systems comprise two main tasks: feature extraction and pattern recognition. The feature extraction stage is designed to transform the incoming speech signal into a relevant representation that serves as the input to a later pattern recognition stage.

Different feature vectors for ASR have been proposed in the literature, but their derivations remain ad hoc. We propose to define a feature vector based on a perceptually relevant objective criterion. The human peripheral auditory system enhances the input speech signal for further processing by the central auditory system of the brain. Pre-processing of the input speech signal by the human auditory periphery forms a useful basis for designing an efficient feature vector. Commonly used feature vectors use knowledge of the auditory system in an *ad hoc* manner (non-optimal manner[1]). For example, several feature vector extraction methods perform auditory frequency filtering on a perceptually motivated frequency scale rather than a linear scale. Another example is the use of a logarithmic function to approximate the nonlinear dynamic compression in the auditory system, which allows coverage of the large dynamic range between hearing threshold and uncomfortable loudness level. Using these two auditory motivated signal processing techniques, for example, mel frequency cepstral coefficients (MFCCs) were designed a few decades ago [1], [2]. They are still universally used due to their computational simplicity as well as their good performance. We note that the MFCCs do not use up-to-date quantitative knowledge of the auditory system.

Several attempts have been made to use quantitative auditory models in a practical ASR system processing chain [3]–[10]. In these techniques, the input speech signal is first processed through a readily available auditory model and then the output signal of the auditory model is formatted to use as an input feature vector to the pattern recognition stage of the ASR system. For example, in [7], it was shown that a feature vector derived from the direct use of a spectro-temporal auditory model (proposed in [14], [15]) provides better performance than the MFCCs. The direct use of an auditory model was shown to provide better speech recognition performance, but at the expense of higher computational complexity. In recent years, the research in quantitative modeling of the complex peripheral auditory system has reached a high level of sophistication [11]–[18], and it is appealing to use a sophisticated auditory model for designing efficient feature vectors. Also, there have been attempts where the parameters of a feature vector is optimized by considering the feedback from the ASR system, for example, using a discriminant analysis based approach in [19]. In this paper, we endeavor to design sophisticated auditory model based feature vectors with the attribute of computational simplicity. Also, we endeavor to design feature vectors only using auditory motivated knowledge, but without any feedback from the ASR system.

For ASR tasks, most of speech feature vectors (for example, MFCCs) are computed using a short term speech analysis technique where spectral content stationarity is assumed over a short segment of speech signal. Based on a short term speech analysis technique, the vector of MFCCs can be regarded as a static spectral feature vector that is computed frame-by-frame independently. To use spectro-temporal properties of speech signal in an ASR system, the standard practice is to use the spectro-temporal dynamic feature vectors (such as velocity and acceleration vectors), which are computed from the static feature vector

[1] Non-optimal in the sense that the parameters of a feature vector is not optimized for an auditory-motivated objective measure.

Manuscript received April 19, 2010; revised August 07, 2010; accepted December 07, 2010. Date of publication December 23, 2010; date of current version June 01, 2011. This work was supported by the EU FP6 ACORNS project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

S. Chatterjee is with the Communication Theory Lab, School of Electrical Engineering, KTH-Royal Institute of Technology, 10044 Stockholm, Sweden (e-mail: saikatchatt@gmail.com; sach@kth.se).

W. B. Kleijn is with the School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6012, New Zealand (e-mail: bastiaan.kleijn@ecs.vuw.ac.nz).

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.2010.2101597

using a standard regression method. The dynamic feature vectors are concatenated with the static feature vector to construct a combined feature vector. Even though the velocity and acceleration feature vectors help to improve ASR performance, they are ad hoc in nature and do not use the up-to-date knowledge of spectro-temporal auditory properties. In the literature, we note that considerable attention is paid to the development of efficient static feature vectors, but not to efficient dynamic feature vectors. A standard regression method remains as a predominantly used and computationally simple *de-facto* method to obtain a dynamic feature vector from a static feature vector.

In this paper, we propose a perturbation theory based offline optimization framework for designing improved feature vectors using state-of-the-art auditory models. Using the framework, we address two issues: 1) optimizing a static spectral feature vector using a spectral auditory model to incorporate auditory response across frequency, and 2) designing and optimizing a new dynamic spectro-temporal feature vector using a spectro-temporal auditory model to incorporate auditory response across time and frequency.

We investigate the use of both spectral and spectro-temporal auditory models to design improved static and dynamic feature vectors through offline optimization instead of the direct online use. The offline optimization approach of using complex auditory models helps to retain the computational design simplicity of the feature vectors. Also, it avoids the difficulty of formatting the output of an auditory model for recognition. In our approach, a feature vector is optimized in such a way that it emulates the behavior of the human auditory system. The implementation of our method relies on perturbation theory and does not consider any feedback from the ASR system.

Our objective is to come close to the situations where (speech) classification in the feature domain is identical to classification in the auditory domain based on the assumption that the auditory domain is good for speech classification. In general, the classification of data in two different domains will provide the same results if the geometry of the space is identical. If the geometry is identical, then the mapping between the spaces is an isometry, or distance preserving. This implies that the norms, including the Euclidean distances, are preserved. Most important in preserving classification across two domains is the "local" geometry. A "slow" global warping would not affect classification and since we do not want to be restrictive, we only check the local geometry (i.e., we check only small distances). Thus, we conjecture that human-like classification of speech sounds is facilitated by similarity between the local geometries of two domains, the feature vector domain and the auditory domain. The basis of the conjecture is that the preservation of the data geometry near the class boundaries is most critical for improved classification. This means that a "small" Euclidean distance in the auditory domain must be similar to a "small" Euclidean distance in the feature domain, except for an overall scaling. The focus on small distances allows a complex perceptual distance to be reduced to a quadratic distance measure using a sensitivity matrix based analysis. The sensitivity matrix-based analysis was first developed in the context of source coding [20]. In [21], the sensitivity matrix was used to simplify an auditory distance measure for audio coding. Here, we extend the sensitivity matrix paradigm to optimize a feature vector. Through experimental evaluation, we support the conjecture based optimization approach.

In [22] and [23], for dimension reduction of feature vectors in ASR, the approach of preserving local geometry using sensitivity matrix-based analysis was shown to be better than the commonly used discriminant analysis-based methods, such as linear discriminant analysis (LDA) and heteroscedastic LDA (HLDA). Note that the approach is fully based on auditory knowledge and does not require classified data (labeled training data) as used in the discriminant analysis-based methods (therefore, no feedback from the ASR system is required). In this paper, our objective is to improve a feature vector, but not to address the problem of dimension reduction. We introduce some adjustable parameters in a feature vector and optimize the parameters such that the similarity between the local geometries improves.

For optimizing a spectral feature vector, we apply the perturbation theory-based optimization framework towards improvement of standard MFCCs using a spectral auditory model. We use the MFCC feature vector as the base feature vector for further improvement. The choice of the MFCCs is based on their wide-spread use, computational simplicity and suitable design architecture. The optimized MFCCs are referred to as *modified MFCCs* (mMFCCs) [24]. Comparing to traditional MFCCs, the mMFCCs have a similar structure as well as computational simplicity. Using the standard HTK (hidden Markov model (HMM)-based toolkit) software [25], the optimized mMFCCs are shown to provide better ASR performance than the standard MFCCs for both clean and noisy acoustic conditions. Next, we consider to design and optimize a new dynamic feature vector using a spectro-temporal auditory model to incorporate auditory response across time and frequency. The development of the dynamic feature vector is based on the evaluation architecture of standard MFCCs and hence retains the computational simplicity. Further improvement in ASR performance is obtained when the dynamic feature vector is used along-with the standard mMFCC feature vector.

II. Maximizing Similarity Between Spaces

Improvement in sound classification requires a feature vector representation that provides a good separation of sound classes in the feature vector space. Therefore, we optimize a feature vector to better describe the inter-sound distances of a state-of-the-art auditory model. We conjecture that if the Euclidean distance between two acoustic-feature vectors approximates the corresponding perceptual distortion for two different speech sounds, then the use of that acoustic feature vector generally leads to better classification in an ASR system. Ideally, this implies an isometry between the perceptual and feature domains. The mapping from the perceptual to feature domain would then be *distance preserving*.

A. Distance Preserving Measure

In practice, it may not be possible to design a feature vector that leads to an accurate distance-preserving mapping from perceptual domain to feature domain. However, it is not required to preserve all the distances. For good classification, the preservation of the data geometry near the class boundaries is most critical. More generally, the preservation of small distances (reflecting the local geometry) near the classification boundary is important, whereas the preservation of large distances (reflecting the global geometry) is not required. In principle, to achieve better sound classification, we then simply desire to have the same small distances for the auditory domain and for the feature domain.

A feature vector is a function of an input speech signal segment (or speech frame) and some adjustable design parameters. For example, to design mMFCCs, these design parameters can be the frequency warping parameter to change the shape of a filter bank (such as gains, bandwidths, center frequency of filters), a parameter to change the shape of a compressing function (like logarithmic function), etc. The objective is to obtain a feature vector with optimum parameters for which any small perturbation of the input speech signal segment leads to a Euclidean distance in the feature domain that best approximates the perceptual distortion indicated by the auditory model. Naturally, this criterion has to hold for all speech segments. To measure the similarity of the auditory model distortion and the feature domain distance, a suitable objective measure needs to be designed that will provide a means of ensemble averaging over all speech segments and all perturbations. By optimizing the parameters, a higher similarity in the objective measure corresponds to a better feature vector.

We now define an objective measure that relates between the perceptual and feature domains. We denote the signal vector for the jth speech frame as $\mathbf{x}_j \in \mathbb{R}^N$, where $j \in J \subset \mathbb{Z}$. Note that, depending upon an application requirement of using either a spectro-temporal auditory model or a spectral auditory model, we treat the signal vector $\mathbf{x}_j$ either as a time domain signal or as a power spectrum domain signal. Let us denote the perceptual domain representation of $\mathbf{x}_j$ as $\mathbf{y}: \mathbb{R}^N \rightarrow \mathbb{R}^K$. For a feature vector, we denote the adjustable design parameters by a vector $\mathbf{p} \in \mathbb{R}^S$. Then, we can denote the Q-dimensional feature vector derived from $\mathbf{x}_j$ using $\mathbf{p}$ as $\mathbf{c}: \mathbb{R}^N \times \mathbb{R}^S \rightarrow \mathbb{R}^Q$. The perceptual domain distortion is defined through a mapping as $\phi: \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of non-negative reals. For the jth speech frame, let us denote the lth perturbed signal as $\hat{\mathbf{x}}_{j,l}$. Often the perceptual distortion measure is based on the L^2 norm of the difference between the perceptual domain signal $\mathbf{y}(\mathbf{x}_j)$ and its distorted version $\mathbf{y}(\hat{\mathbf{x}}_{j,l})$. In that case, the perceptual distortion $\phi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}) = ||\mathbf{y}(\mathbf{x}_j) - \mathbf{y}(\hat{\mathbf{x}}_{j,l})||^2$. Using the L^2 norm, we can define a distance measure for the feature vector $\mathbf{c}(\mathbf{x}_j, \mathbf{p})$ as $\psi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p}) = ||\mathbf{c}(\mathbf{x}_j, \mathbf{p}) - \mathbf{c}(\hat{\mathbf{x}}_{j,l}, \mathbf{p})||^2$. Now, considering the finite sequence of speech frames $j \in J$ and a finite set of acoustic perturbations $l \in L_j$, the objective is to minimize a measure of dissimilarity between perceptual domain distortion and feature domain distortion with respect to the adjustable design parameter vector $\mathbf{p}$. To satisfy this objective, a suitable norm based objective measure can be defined as

$$\xi = \sum_{j \in J} \sum_{l \in L_j} \left[\phi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}) - \lambda\ \psi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})\right]^2 \quad (1)$$

where

$$\lambda = \frac{\sum_{j \in J} \sum_{l \in L_j} \phi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l})\ \psi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})}{\sum_{j \in J} \sum_{l \in L_j} \left(\psi(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})\right)^2}. \quad (2)$$

Here λ is the necessary scaling to eliminate the effect of a scale mismatch between perceptual domain and feature domain. So, the objective is to minimize the norm based objective measure ξ with respect to the adjustable parameter vector $\mathbf{p}$.

B. Perturbation Analysis

It may be possible to directly compute complex distortion measures and hence, to minimize the objective measure of (1) even for complex distortion measures. The approach of directly evaluating complex distortion measures is computationally expensive. Also the approach may not render sufficient analytical/algorithmical tractability for optimization. Since we are interested in small distances, we can approximate the perceptual and feature domain distortion measures using simpler quadratic distortion measures, leading to a significant reduction in computational complexity and an increase in analytical/algorithmical tractability. This approach is based on the sensitivity matrix framework [20], [21].

Gardner and Rao [20] were the first to introduce approximation of a distortion measure by means of the sensitivity matrix. Based on this work, Plasberg and Kleijn [21] derived the sensitivity matrix for a complex perceptual distortion. Here, we briefly outline the main concepts of the sensitivity matrix-based distortion measure analysis.

Let us omit the subscripts for notational brevity where no ambiguity exists. We assume that the perceptual domain distortion $\phi(\mathbf{x}, \hat{\mathbf{x}})$ is continuous and its third-order derivatives exist. Also, $\phi(\mathbf{x}, \hat{\mathbf{x}}) \geq 0$ with equality iff $\hat{\mathbf{x}} = \mathbf{x}$. Then, for a sufficiently small perturbation $\hat{\mathbf{x}} - \mathbf{x}$, that is, closely around the unique minimum $\hat{\mathbf{x}} = \mathbf{x}$ of $\phi(\mathbf{x}, \hat{\mathbf{x}})$, the first-order partial derivatives of the Taylor series expansion of $\phi(\mathbf{x}, \hat{\mathbf{x}})$ vanish and we can write

$$\phi(\mathbf{x}, \hat{\mathbf{x}}) \approx \frac{1}{2}[\hat{\mathbf{x}} - \mathbf{x}]^T \mathbf{D}_\phi(\mathbf{x})\, [\hat{\mathbf{x}} - \mathbf{x}] \quad (3)$$

where $\mathbf{D}_\phi(\mathbf{x})$ is the $N \times N$-dimensional positive semidefinite sensitivity matrix whose elements are $\mathbf{D}_{\phi,ij}(\mathbf{x}) = \partial^2 \phi(\mathbf{x}, \hat{\mathbf{x}}) / \partial \hat{x}_i\, \partial \hat{x}_j \big|_{\hat{\mathbf{x}}=\mathbf{x}}$. The effect of the third and higher order derivatives is neglected in (3). In certain cases, such as the spectral and spectro-temporal auditory models of Section III, $\mathbf{D}_\phi(\mathbf{x})$ can be directly evaluated from input signal $\mathbf{x}$ and hence, $\phi(\mathbf{x}, \hat{\mathbf{x}})$ can be evaluated for small perturbation $\hat{\mathbf{x}} - \mathbf{x}$. Note that $\hat{\mathbf{x}} - \mathbf{x}$ is the only term in (3) that depends on $\hat{\mathbf{x}}$, which explains the reduction in computational complexity for repeated evaluation of perceptual distortion $\phi(\mathbf{x}, \hat{\mathbf{x}})$.

Next, we consider a simplification of the feature domain distortion defined as $\psi(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{p}) = ||\mathbf{c}(\mathbf{x}, \mathbf{p}) - \mathbf{c}(\hat{\mathbf{x}}, \mathbf{p})||^2$. We assume that the mapping $\mathbf{c}(\mathbf{x}, \mathbf{p})$ is continuous and its second-order derivatives exist. Then, for a sufficiently small perturbation $\hat{\mathbf{x}} - \mathbf{x}$, we use a Taylor series expansion to make a local approximation around $\mathbf{x}$ as

$$\mathbf{c}(\hat{\mathbf{x}}, \mathbf{p}) \approx \mathbf{c}(\mathbf{x}, \mathbf{p}) + \mathbf{A}(\mathbf{x}, \mathbf{p})\ [\hat{\mathbf{x}} - \mathbf{x}] \quad (4)$$

where $\mathbf{A}(\mathbf{x}, \mathbf{p})$ is a $Q \times N$-dimensional matrix as $\mathbf{A}(\mathbf{x}, \mathbf{p}) = \partial \mathbf{c}(\mathbf{x}, \mathbf{p}) / \partial \hat{\mathbf{x}} \big|_{\hat{\mathbf{x}}=\mathbf{x}}$. The effect of the second and higher order derivatives is neglected in (4). Then, we can write the feature domain distortion as

$$\begin{aligned}\psi(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{p}) &= ||\mathbf{c}(\mathbf{x}, \mathbf{p}) - \mathbf{c}(\hat{\mathbf{x}}, \mathbf{p})||^2 \\ &\approx [\hat{\mathbf{x}} - \mathbf{x}]^T \mathbf{A}(\mathbf{x}, \mathbf{p})^T \mathbf{A}(\mathbf{x}, \mathbf{p})\, [\hat{\mathbf{x}} - \mathbf{x}]. \end{aligned} \quad (5)$$

For given $\mathbf{A}(\mathbf{x}, \mathbf{p})$, note that $\hat{\mathbf{x}} - \mathbf{x}$ is the only term in (5) that depends on $\hat{\mathbf{x}}$, which explains the reduction in computational complexity for repeated evaluation of feature domain distortion $\psi(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{p})$.

III. Auditory Models

For designing static and dynamic feature vectors, the corresponding auditory models are briefly discussed below. The models are used to evaluate distances (L^2 norms) between sounds in the auditory domain. Note that for our purpose, the distances do not have to correspond to a quantitative measure of perceived difference between the sounds.

A. Spectral Auditory Model

For optimization of a static feature vector, we use the spectral auditory model developed by van de Par, *et al.* [18] which is referred to as the van de Par auditory model (VAM). In the VAM, each speech frame is independently analyzed and the power spectrum signal of each frame is used as the input signal. For computing the power spectrum using a standard periodogram technique, a hamming windowed speech frame is DFT transformed followed by evaluating the square absolute values of the DFT coefficients. The VAM consists of several frequency channels, in each of which the ratio of distortion power to masker power is calculated. Then, the ratios of all the frequency channels are combined together to account for the spectral integration property of the human auditory system. For the jth speech frame, let the N-dimensional vector $\mathbf{z}_j = [z_{j,0}\ z_{j,1}, \ldots, z_{j,n}, \ldots, z_{j,N-1}]^T$ be the power spectrum of corresponding time domain signal. Let $\mathbf{H}$ be a diagonal N-dimensional matrix whose diagonal is formed by the frequency response of the outer and middle ear filter. In the same fashion, a diagonal $\mathbf{G}_i$ is defined, so that the frequency response of the ith channel Gamma-tone auditory filter forms its diagonal. For the VAM, the diagonal sensitivity matrix of jth frame is

$$\mathbf{D}_\phi(\mathbf{z}_j) \approx 2\frac{C_s L_e}{N} \sum_i \frac{[\mathbf{G}_i \mathbf{H}]^T [\mathbf{G}_i \mathbf{H}]}{\frac{1}{N}[\mathbf{G}_i \mathbf{H} \mathbf{z}_j]^T [\mathbf{G}_i \mathbf{H} \mathbf{z}_j] + C_a} \tag{6}$$

where C_s and C_a are constants calibrated based on measurement data, and L_e is a constant to account for the influence of temporal integration time in the human auditory system on frame duration. Using VAM, the perceptual distortion is expressed as

$$\phi(\mathbf{z}_j, \hat{\mathbf{z}}_{j,l}) \approx \frac{1}{2}[\hat{\mathbf{z}}_{j,l} - \mathbf{z}_j]^T \mathbf{D}_\phi(\mathbf{z}_j)\, [\hat{\mathbf{z}}_{j,l} - \mathbf{z}_j]. \tag{7}$$

Given the VAM sensitivity matrix $\mathbf{D}_\phi(\mathbf{z}_j)$, we note that the perceptual distortion is computed using the power spectrum domain speech signal $\mathbf{z}_j$ and its lth perturbed version $\hat{\mathbf{z}}_{j,l}$.

It is important to mention that each speech frame is independently analyzed in the VAM. Therefore, the use of the VAM is appropriate for optimizing a static feature vector. Note that due to the inability to model the auditory response in time, the use of the VAM is inappropriate for optimizing a temporal dynamic feature vector.

B. Spectro-Temporal Auditory Model

For design and optimization of a dynamic spectro-temporal feature vector, we use the spectral-temporal auditory model developed by Dau *et al.* [14], [15]. This spectro-temporal auditory model is referred to as the Dau auditory model (DAM). The DAM quantifies and transforms an incoming sound waveform into its “internal” representation [14]. A block diagram of the DAM is shown in Fig. 1 [7], [14]. To model the basilar membrane, the input speech signal is decomposed into the critical band signals using a gamma-tone filterbank. Next, representing a hair-cell model in each critical band channel, the output signal of each gamma-tone filter is half-wave rectified and first-order low-pass filtered (with a cutoff frequency of 1 kHz) for envelope extraction. Therefore, at this processing stage, each critical band frequency channel contains information about the amplitude variation of the input signal within the channel. This envelope is then compressed using an adaptive circuit consisting of five consecutive *nonlinear* adaptive compression loops (ACLs). Each of these loops consists of a divider and a first order IIR low-pass filter (LPF). The time constants of LPFs of five loops are: $\tau^{(1)} = 5$ ms, $\tau^{(2)} = 50$ ms, $\tau^{(3)} = 129$ ms, $\tau^{(4)} = 253$ ms, and $\tau^{(5)} = 500$ ms (the cutoff frequencies are 32, 3.2, 1.23, 0.62, and 0.32 Hz, respectively). For each adaptive loop, the input signal is divided by the output signal of the low-pass filter. Sudden transitions in a critical band envelope that are fast compared to the time constants of the ACLs are amplified linearly at the output due to slow changes in the outputs of low-pass filters, whereas the slowly changing portions of the envelope are compressed. Due to this transformation characteristic, changes in the input signal like onsets and offsets are emphasized, whereas the steady state portions are compressed. The nonlinear ACLs function introduces long inherent memory in the model and help to take into account the dynamic temporal structure of auditory response. The last processing step is a first-order LPF with a cutoff frequency of 8 Hz to optimize predictions of psycho-acoustical masking experiments. This LPF acts as a modulation filter and attenuates fast envelope fluctuations of the signal in each critical band frequency channel.

The ability of the original DAM to describe data from modulation-detection and modulation-masking experiments was further examined in [16]. The model of [16] was built on the original DAM of [14] and only a single substantial change was introduced while keeping the original DAM intact. In the last processing step of each critical band channel, instead of the LPF with a cutoff frequency of 8 Hz, a modulation filterbank was introduced to analyze the amplitude changes of the envelope. The nonuniform filters of the modulation filter bank follows a logarithmic scaling and the lowest modulation filter is a LPF with a cutoff frequency of 2.5 Hz [16].

For the optimization of a dynamic feature vector, we use the perturbation-based framework where the perceptual distortion is computed through the use of a sensitivity matrix. The issue of evaluating a sensitivity matrix for the original DAM was recently addressed in [21]. To develop an optimized dynamic feature vector, we use the original DAM and its readily available sensitivity matrix solution of [21].

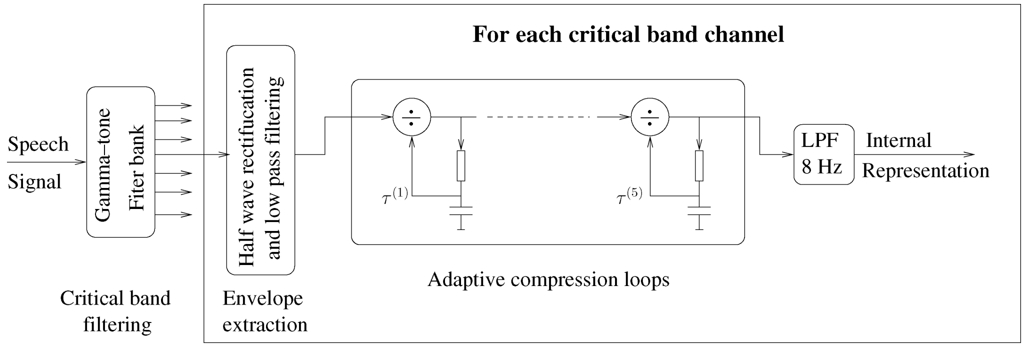

Fig. 1. Spectro-temporal auditory model: Dau auditory model ("DAM").

For the DAM, the derivation of the sensitivity matrix is tedious. The details of evaluating the DAM sensitivity matrix are described in [21]. Let N-dimensional vector $\mathbf{s}_j = [s_{j,0}\ s_{j,1}, \ldots, s_{j,n}, \ldots, s_{j,N-1}]^T$ denotes time domain speech signal for the jth speech frame. Using $\mathbf{s}_j$, the perceptual distortion is expressed as

$$\phi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}) \approx \frac{1}{2}[\hat{\mathbf{s}}_{j,l} - \mathbf{s}_j]^T \mathbf{D}_\phi(\mathbf{s}_j)\, [\hat{\mathbf{s}}_{j,l} - \mathbf{s}_j]. \tag{8}$$

Given the DAM sensitivity matrix $\mathbf{D}_\phi(\mathbf{s}_j)$, it can be noted that the perceptual distortion is computed using the time domain speech signal $\mathbf{s}_j$ and its lth perturbed version $\hat{\mathbf{s}}_{j,l}$.

IV. Spectral Auditory Model-Based Static Feature Vector

Conventional speech processing techniques use a short-term speech analysis technique where the input speech signal is framed into short segments (typically 20–40 ms) with a reasonable frame shift (typically 10 ms). Most speech feature vectors, like MFCCs, are evaluated from the power spectrum signal computed from a short time speech segment. To optimize a power spectrum based feature vector, we use the spectral auditory model VAM.

We first generalize the definition of the MFCCs to render a feature vector with adjustable design parameters $\mathbf{p}$. We refer to this new feature vector as *modified MFCCs* (mMFCCs) [24]. Using the power spectrum $\mathbf{z}_j$ as an input, the steps of evaluating the mMFCCs for jth speech frame are as follows.

1) Calculation of filter bank energies (FBEs): The energy in each frequency channel is computed as

$$\begin{aligned} e_{j,m} &= \mathbf{z}_j^T\ \mathbf{w}_m(\alpha) \\ &= \sum_{n=0}^{N-1} z_{j,n} \times w_{m,n}(\alpha)\,, 0 \le m \le M-1 \end{aligned} \tag{9}$$

where $\mathbf{w}_m(\alpha)$ is the N-dimensional vector denoting the triangular filter of the mth channel and satisfies $\sum_{n=0}^{N-1} w_{m,n}(\alpha) = 1$. M is the total number of channels with a typical value of $M = 26$. The triangular filters are equispaced and linearly arranged along the perceptually motivated warped frequency scale. Therefore, the triangular filters are placed in a nonuniform manner along the normal frequency scale. The shape of a triangular filter depends on the extent of frequency warping. We generalize the relation between the normal frequency scale and the *mel* frequency scale [2] and define a perceptually motivated warped frequency scale

$$f_{\text{warp}} = 2595 \times \log_{10}\left(1 + \left(\frac{f_{\text{norm}}}{\alpha}\right)\right) \tag{10}$$

where α is the warping factor and f_{norm} is the normal frequency scale in Hz. An increase in α leads to a decrease in the extent of warping. For the mMFCCs, α is a parameter to optimize to achieve better recognition performance. In the case of standard MFCCs, the triangular filters are designed using the *mel* frequency scale where $\alpha = 700$ [2], [1].

2) Static compression: Compression of the dynamic range of FBEs using polynomial logarithmic function as follows:

$$\gamma_{j,m} = \log_{10}\left[\sum_{r=1}^{R} b_r\,(e_{j,m})^r\right],\ 0 \le m \le M-1 \tag{11}$$

where $\sum_{r=1}^{R} b_r = 1$ and $b_r \ge 0$. For the mMFCCs, we optimize the polynomial coefficients $\{b_r\}_{r=1}^{R}$. In the case of standard MFCCs, we use the logarithmic function, i.e., $R = 1$ and $b_1 = 1$ [2], [1]. We note that (11) implies that our results are scale dependent and require proper normalization.

3) De-correlation: De-correlation using the DCT to evaluate Q-dimensional mMFCC feature vector $\mathbf{g}_j$ whose elements are

$$g_{j,q} = \sum_{m=0}^{M-1} \gamma_{j,m} \times \cos\left[q\,(m + 0.5)\,\frac{\pi}{M}\right],\ 1 \le q \le Q. \tag{12}$$

A typical value of feature vector dimension is $Q = 12$.

A. Optimization of mMFCCs

The adjustable design parameters that we optimize to obtain the mMFCCs are $\mathbf{p} = \left[\alpha, \{b_r\}_{r=1}^{R}\right]$. To optimize the adjustable design parameters, we minimize the objective measure ξ of (1) where the input signal is the power spectrum domain signal $\mathbf{z}_j$ [for mMFCCs, we treat $\mathbf{x}_j = \mathbf{z}_j$ in (1)]. In this case, we note that the objective measure of (1) is a function of the perceptual distortion $\phi(\mathbf{z}_j, \hat{\mathbf{z}}_{j,l})$ and the feature domain distortion $\psi(\mathbf{z}_j, \hat{\mathbf{z}}_{j,l}, \mathbf{p})$. Using the VAM, the perceptual distortion in spectral domain is shown in (7). The feature domain distortion $\psi(\mathbf{z}_j, \hat{\mathbf{z}}_{j,l}, \mathbf{p})$ is shown in (5) where the input signal is the power spectrum domain signal $\mathbf{z}_j$. For the feature domain distortion $\psi(\mathbf{z}_j, \hat{\mathbf{z}}_{j,l}, \mathbf{p})$, the $Q \times N$-dimensional matrix $\mathbf{A}(\mathbf{z}_j, \mathbf{p}) = [A(\mathbf{z}_j, \mathbf{p})_{q\,n}]_{Q \times N}$ has the elements

$$\begin{aligned} A(\mathbf{z}_j, \mathbf{p})_{q\,n} &= \frac{\partial g_{j,q}}{\partial z_{j,n}} = \frac{\partial g_{j,q}}{\partial \gamma_{j,m}} \frac{\partial \gamma_{j,m}}{\partial e_{j,m}} \frac{\partial e_{j,m}}{\partial z_{j,n}} \\ &= \sum_{m=0}^{M-1} \cos\left[q\,(m+0.5)\,\frac{\pi}{M}\right] \\ &\quad \times \frac{\sum_{r=1}^{R} r\, b_r\, (e_{j,m})^{r-1}}{\ln 10 \times \sum_{r=1}^{R} b_r\, (e_{j,m})^{r}}\, w_{m,n}(\alpha). \qquad (13) \end{aligned}$$

It is interesting to jointly optimize all the parameters through a closed-form/iterative solution, such as a gradient descent search technique. To minimize the objective measure ξ of (1), a gradient descent search technique requires a closed form gradient expression of $d\xi/d\mathbf{p}$, which is not easy to evaluate due to the intricate relationship existing between the objective measure ξ of (1) and the parameter vector $\mathbf{p} = \left[\alpha, \{b_r\}_{r=1}^{R}\right]$. Therefore, we use an iterative optimization algorithm, based on an increment-based search, to jointly optimize the parameters through the minimization of the objective measure ξ. We start the algorithm using an initial value of $\alpha = 700$ (standard MFCCs use $\alpha = 700$). For a given α, we optimize $\{b_r\}_{r=1}^{R}$ through the minimization of ξ using an increment-based search. Then, for given optimized $\{b_r\}_{r=1}^{R}$, we optimize α through the minimization of ξ using an increment-based search. Then we iteratively repeat the increment-based search process of optimizing $\{b_r\}_{r=1}^{R}$ using the previous optimized α and α using the previous optimized $\{b_r\}_{r=1}^{R}$. This process of optimization iteratively goes on until the objective measure ξ of (1) reaches a local minima with the optimum values of α and $\{b_r\}_{r=1}^{R}$.

For both the cases of wideband (sampling frequency of 16 kHz) and narrowband (sampling frequency of 8 kHz) speech, we use five minutes of clean speech training data (after removing silence region) to optimize the adjustable parameters of mMFCCs. To evaluate the mMFCCs, we use a 32-ms Hamming windowed speech frame with 10-ms frame shift and the number of triangular filters $M = 26$ and the mMFCC feature vector dimension $Q = 12$. The power spectrum of each frame is computed using a standard DFT-based periodogram technique and the power spectrum is perturbed with independent and identically distributed (i.i.d.) Gaussian noise at signal-to-noise ratios (SNRs) ranging from 40 to 50 dB with an increment step of 0.1 dB. In the perturbation based approach, it is important to note the range of validity for the linearization assumption. Using scatter plots, we can compare the changes in the feature vectors computed from the linearized relation [see (5) and (13)] with the true difference between the feature vectors of original and distorted power spectrum signals. A high correlation in scatter plots allows us to assume the validity of linearization. Using the iterative optimization technique, we find that a polynomial order of $R = 2$ is sufficient[2]; the optimal values of the polynomial coefficients are $b_1 = 0.1$ and $b_2 = 0.9$. For wideband speech and narrowband speech, we find the optimum values of the warping factor are $\alpha = 900$ and $\alpha = 1100$, respectively. We note that standard MFCCs use $b_1 = 1$ and $\alpha = 700$ irrespective of the sampling frequency of the input speech, the choice of the window length and the window shift, and the feature vector dimension (Q) and the number of frequency channels (M) [1], [2].

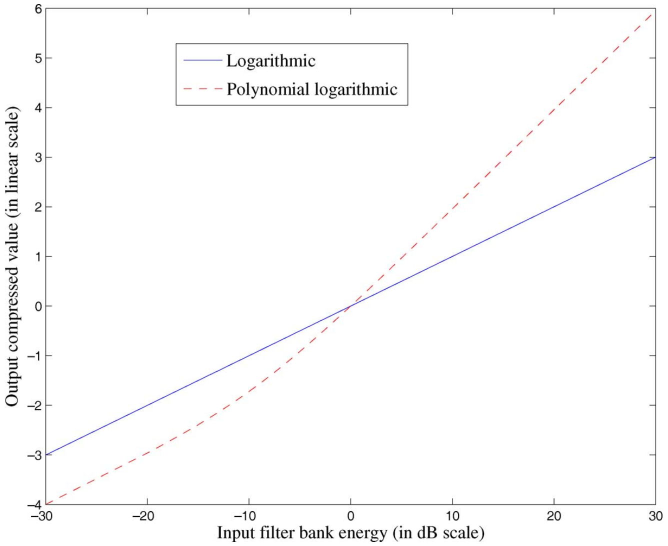

Fig. 2. Input–output behaviors of compression functions: logarithmic and polynomial logarithmic.

Using our perturbation-based sensitivity analysis approach and generalizing the definition of the standard MFCCs, we are able to determine the optimal warping, order and coefficients of the optimum polynomial logarithmic function. For mMFCCs, we show the optimal polynomial logarithmic compressing function in Fig. 2 and compare it with the standard logarithmic compressing function as used in the standard MFCCs. It is observed that a higher input filter bank energy level is emphasized more than a lower input level. This trend of the input–output transfer function is some extent similar with the power law nonlinearities as described by Steven's power law of hearing [30] and used in designing other perceptually motivated feature vectors, such as the perceptual linear prediction coefficients [31] and the recently proposed power-normalized cepstral coefficients [32]. For example, the perceptual linear prediction (PLP) coefficients [31] use a power law nonlinearity with the exponent of 0.33 based on the Steven's power law of hearing. Let us discuss the input–output transfer function for the case that where we use the power law nonlinearity (with the exponent of 0.33) to compress the FBEs. Compared to the polynomial logarithmic function, we find that the use of the power law nonlinearity function (with the exponent of 0.33) leads to more emphasis on a higher input FBE signal level and much less on a lower input FBE signal level. Due to such a behavior, the discriminative power in the output level for the lower FBE signal level region is severely affected in the case of using the power law nonlinearity. Through informal

[2] Using a third-order polynomial, i.e., using $R = 3$, we found that the coefficient b_3 is nearly zero (very small). Therefore, we decided to use a second-order polynomial.

ASR experiments, we noted that the direct use of the power law nonlinearity, instead of the polynomial logarithmic function, to compress the FBEs could not lead to a better feature vector.

V. Spectro-Temporal Auditory Model-Based Dynamic Feature Vector

In this section, we develop a computationally simple dynamic feature vector that uses both spectral and temporal auditory properties. For designing of such a dynamic feature vector, we use the dynamic spectro-temporal auditory model DAM.

The use of the DAM has been investigated in the literature for designing suitable feature vectors. In [7], the output of the DAM was formatted and directly used as a feature vector in an ASR system. For speaker-independent digit recognition, the DAM based feature vector was shown to better than the MFCCs (see [7, Fig. 6]). Recently, the use of ACLs of the DAM was investigated for deriving modulation-based feature vector in [27], where the input speech signal was decomposed into critical band signals and then the temporal envelopes of sub-band signals were compressed using static (logarithmic) and adaptive (ACLs) compressions separately. The feature vector of [27] was shown to be better than the MFCCs. However, the feature vector of [27] is computationally intensive and also, the feature vector is of high dimension (the feature vector dimension is 476 using 17 critical bands and 28 modulation components per sub-band). To design a computationally simple feature vector with moderate dimensionality, we investigate the use of nonlinear ACLs followed by a LPF for compressing the FBEs of mMFCCs alongwith the static polynomial logarithmic compression.

A. *Adaptive Compression-Based Dynamic Feature Vector*

Most speech feature vectors, such as MFCCs and mMFCCs, are static in nature and evaluated in the time–frequency-based analysis framework. On the other hand, a sophisticated auditory model, such as the DAM, uses an alternate framework of frequency–time analysis where the time domain envelope of the output of a gamma-tone filter (frequency channel) is processed through ACLs to incorporate a dynamic auditory response. Therefore, it is interesting to see how the DAM can be used for time–frequency analysis based feature designn.

To design and optimize a new dynamic feature vector, we use the nonlinear ACLs followed by a modulation LPF to adaptively compress and filter the FBEs across speech frames. The FBEs are already computed in the static mMFCC feature vector evaluation as shown in (9). For the mth triangular filter, the FBE signal $e_{j,m}$ can be viewed as a time domain signal where the index j denotes the time variable. In case of a typical frame shift of 10 ms, the $e_{j,m}$ signal is sampled at a rate of 100 Hz. For the mth triangular filter, the time domain $e_{j,m}$ signal is passed through the ACLs followed by a modulation LPF to incorporate the auditory response across speech frames into the new dynamic feature vector. For the jth speech frame, the steps to compute the new dynamic feature vector are as follows:

1) Adaptive compression: For the mth triangular filter, adaptive compression of $e_{j,m}$ signal as

$$\rho_{j,m} = \mathrm{acl}(e_{j,m}^{\kappa}),\ 0 \leq m \leq M-1 \tag{14}$$

where $\mathrm{acl}(.)$ is the model function of the nonlinear ACLs as used in the DAM and κ is introduced to make the argument representation as close to the envelope of the output of the corresponding gamma-tone filter of the DAM, albeit at a much lower rate of 100 Hz.

2) Modulation filtering: For the mth triangular filter, the signal $\rho_{j,m}$ is passed through a first order IIR low-pass filter to produce

$$u_{j,m} = v_j * \rho_{j,m},\ 0 \leq m \leq M-1. \tag{15}$$

Here, v_j is the impulse response of the first-order IIR filter. We note that the frequency response of the IIR filter depends on its cutoff frequency f_c.

3) De-correlation: De-correlation using the DCT to evaluate a $\mathcal{Q}$-dimensional new dynamic feature vector $\mathbf{d}_j$ whose elements are

$$d_{j,q} = \sum_{m=0}^{M-1} u_{j,m} \times \cos\left[q\,(m+0.5)\,\frac{\pi}{M}\right],\ 1 \leq q \leq \mathcal{Q}. \tag{16}$$

Like the static part evaluation (i.e., for mMFCCs), the DCT is applied across the triangular filters (frequency channels) and we choose the dimensionality[3] as $\mathcal{Q} = 12$.

The new dynamic feature vector $\mathbf{d}_j = [d_{j,1}\ d_{j,2}, \ldots, d_{j,\mathcal{Q}}]^T$ is referred to as the adaptive compression-based dynamic coefficients (ACDCs). Note that the $\mathrm{acl}(.)$ function has long time constants. It means that the current processing depends on its long memory. We typically address the issue of long memory by starting analysis as far as possible in the silence region preceding the speech signal[4].

In the case of the DAM, we note that the nonlinear ACLs followed by a modulation LPF is used to compress and filter the time domain envelope signal in a critical band channel. On the other hand, for designing ACDCs, the nonlinear ACLs followed by a modulation LPF is used to compress the FBE signal in a triangular filter channel. The sampling rate of the time domain envelope signal in a critical band channel is the same as of the input speech signal (typically 8 or 16 KHz), whereas the FBE signal in a triangular filter channel is sampled at a typical rate of 100 Hz. We note that the ACDCs are computationally inexpensive to evaluate as the ACLs is used on a signal with lower sampling rate.

For the mth triangular filter, the system implementation of ACLs followed by a modulation LPF is illustrated in Fig. 3. We consider that the input signal $e_{j,m}$ is sampled at a typical rate of 100 Hz (for the standard frame shift of 10 ms). The input signal $e_{j,m}$ is first thresholded, modeling the absolute hearing threshold in quiet, and then processed through the chain of five adaptation loops. The threshold $t_{\min}$ limits the dynamic range of the input signal $e_{j,m}$ to 100 dB. In Fig. 3, the adaptation loops only differ in the time constants ($\tau^{(k)}$s) of the LPFs and the thresholds $t_{\min}^{(k)}$, $k \in \{1,2,3,4,5\}$. We discuss the kth adaptive loop in the chain of five adaptation loops. Let us consider that the positive input signal to the kth adaptive loop is denoted by $\tilde{e}_{j,m}^{(k)}$.

[3] Through informal experiments, we have chosen the dimension.

[4] For ASR experiments, utterance-by-utterance analysis is performed in practice. For each utterance, we assume the presence of at-least 500-ms silence region preceeding the speech signal.

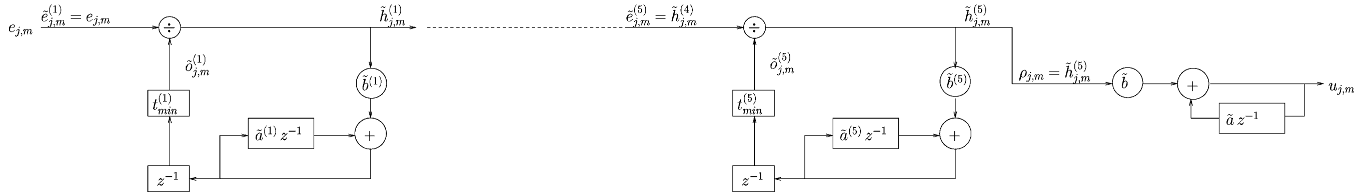

Fig. 3. System implementation of the adaptive compression and modulation filtering of the filter bank energy (FBE) signal for the mth triangular filter (ACLs followed by a modulation LPF).

The signal $\tilde{e}^{(k)}_{j,m}$ is divided by the low-pass filtered version $\tilde{o}^{(k)}_{j,m}$ of the output signal $\tilde{h}^{(k)}_{j,m}$. The LPF is a first-order IIR recursive filter with coefficients $\tilde{a}^{(k)} = \exp(-1/(100 \times \tau^{(k)}))$, where $\tau^{(k)}$ is in the unit of second[5] and $\tilde{b}^{(k)} = 1 - \tilde{a}^{(k)}$. To avoid a division by zero, the denominator $\tilde{o}^{(k)}_{j,m}$ is limited by a threshold $t^{(k)}_{\min} = (t_{\min})^{-2^k}$. For the first adaptive loop, the input signal is the filter bank energy signal, i.e., $\tilde{e}^{(1)}_{j,m} = e_{j,m}$. The output of the fifth adaptive loop is regarded as the output of the ACLs, i.e., $\rho_{j,m} = \tilde{h}^{(5)}_{j,m}$. The output of the ACLs is used as the input signal to the modulation LPF. The modulation LPF is a first-order IIR recursive filter with coefficients $\tilde{a} = \exp(-2\pi f_c/100)$ and $\tilde{b} = 1 - \tilde{a}$; here f_c is the cutoff frequency (in Hz) of the modulation LPF. For the mth triangular filter, the output signal of the modulation LPF is denoted by $u_{j,m}$. After evaluation of $u_{j,m}$ signals for all the triangular filters, the DCT is applied across the triangular filters (i.e., across all channels m, $0 \leq m \leq M-1$) to evaluate the ACDC feature vector.

B. Optimization of ACDCs

The parameters that we optimize to obtain the ACDCs are $\mathbf{p} = [\kappa,\ f_c]$. To optimize the parameters, we minimize the objective measure of (1) where the input signal is the time domain signal $\mathbf{s}_j$ [for ACDCs, we treat $\mathbf{x}_j = \mathbf{s}_j$ in (1)]. In this case, we note that the objective measure of (1) is a function of the perceptual distortion $\phi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l})$ and the feature domain distortion $\psi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}, \mathbf{p})$. Using the DAM, the perceptual distortion $\phi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l})$ in terms of time domain signals is shown in (8). Any feature vector, including ACDCs, is evaluated from the input time domain signal and hence, it is appropriate to denote the feature domain distortion as $\psi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}, \mathbf{p})$. A point to note is that a simple form of feature domain distortion $\psi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}, \mathbf{p})$, using a Taylor series expansion, is not possible to compute in this case due to the following reason. The ACDCs are evaluated by using ACLs on a functional representation of FBEs followed by low-pass filtering and DCT. The FBEs are evaluated from power spectrum signal. Following computationally efficient standard approach, let us decide that the time domain signal $\mathbf{s}_j$ is Hamming windowed and then the power spectrum signal $\mathbf{z}_j$ is computed using a standard DFT-based periodogram technique. The power spectrum signal is a function of input time domain signal. Using the periodogram technique, the computation of power spectrum requires the evaluation of the absolute values of the complex DFT coefficients and hence, the power spectrum is not differentiable with respect to the input time domain signal.[6] Therefore, it is not possible to find a simple form of $\psi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}, \mathbf{p})$ using a Taylor series expansion for the ACDCs. However, for the jth frame, the feature vector is perturbed if the time domain signal is perturbed. Therefore, in this case, we compute the feature domain distortion $\psi(\mathbf{s}_j, \hat{\mathbf{s}}_{j,l}, \mathbf{p})$ directly by computing the L^2 norm of the difference between original ACDC feature vector and perturbed ACDC feature vector.

Like the case of mMFCCs, we use an iterative optimization method and jointly optimize the adjustable design parameters κ and f_c through the minimization of objective measure ξ of (1). For optimization, we use the same five minutes of clean speech training data which was used for optimizing the mMFCCs. Using an initial value of $f_c = 8$ Hz (as used in the DAM), we start the optimization algorithm. We optimize κ for given f_c, and then optimize f_c using the optimized κ. The iteration goes on to minimize the objective measure ξ of (1) and we find the optimum values of κ and f_c at the end.

For both the cases of wideband (sampling frequency 16 kHz) and narrowband (sampling frequency 8 kHz) speech, we use a 32-ms Hamming windowed speech frame with 10-ms frame shift. To evaluate the ACDCs, we use $M = 26$ and $Q = 12$. The time-domain speech signal of each frame is perturbed with i.i.d. Gaussian noise at different SNRs ranging from 40 to 50 dB. We compute a set of perceptual distortions and feature domain distortions using a set of perturbed time domain signals. For both the cases of wideband and narrowband speech, we find the optimum values of the adjustable parameters as $f_c = 4$ Hz and $\kappa = 0.5$ through the use of the iterative optimization technique.

C. Effect of Adaptive Compression on FBEs

The static mMFCCs are evaluated by using a static polynomial logarithmic compression on the FBEs. On the other hand, the adaptive compression-based ACLs function is used to compress the FBEs for evaluating the dynamic ACDCs. Fig. 4 illustrates the effects of static polynomial logarithmic compression and adaptive ACLs' compression on FBEs. A segment of a 1-s speech signal from a long utterance wideband speech signal (16-kHz sampling frequency) is shown in Fig. 4(a). Using 32-ms frame length and 10-ms frame shift, the trajectory of FBE signal for the sixth triangular filter over the frame indices (i.e., $e_{j,6}$) is

[5] Note that, in the unit of second, $\tau^{(1)} = 0.005$ s, $\tau^{(2)} = 0.050$ s, $\tau^{(3)} = 0.129$ s, $\tau^{(4)} = 0.253$ s, and $\tau^{(5)} = 0.500$ s.

[6] Let us consider that a complex operation on a real variable is denoted through a mapping $z : \mathbb{R} \rightarrow \mathbb{C}$. Suppose s is a real variable and hence $z(s)$ is a complex variable. We desire to evaluate the derivative of the absolute value of the complex variable with respect to the real variable, i.e., $\partial|z(s)|/\partial s$. For evaluation, we need to express $\partial|z(s)|/\partial s = \partial|z(s)|/\partial z(s) \times \partial z(s)/\partial s$. For a complex variable $z(s)$, $\partial|z(s)|/\partial z(s)$ does not exist except the point $z(s) = 0$. The derivative does not exist because at every point in the complex plane, the value of the derivative depends on the direction in which it is taken.

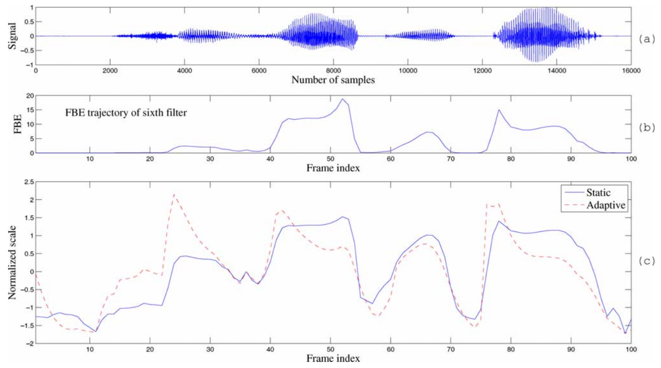

Fig. 4. (a) Segment of 1-s wideband speech signal (16-kHz sampling frequency). (b) The trajectory of FBE signal for sixth triangular filter. (c) Outputs of static (polynomial logarithmic) and adaptive (ACLs) compressions in a normalized scale.

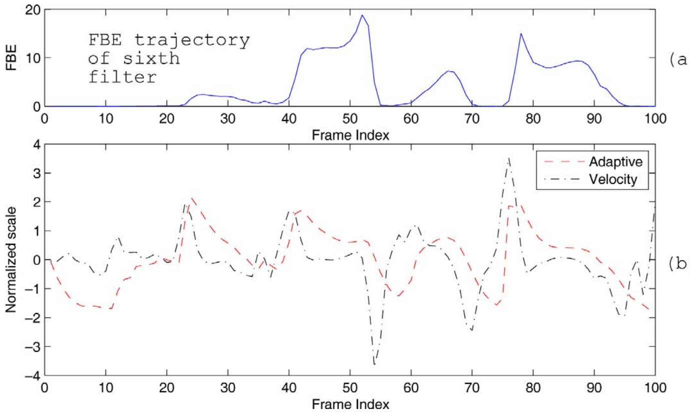

Fig. 5. (a) Trajectory of FBE signal for sixth triangular filter as shown in Fig. 4(b). (b) In a normalized scale: outputs of adaptive compression (ACLs) on FBEs and velocity operation on statically (polynomial logarithmic) compressed FBEs.

shown in Fig. 4(b). Fig. 4(c) shows the results of static compression (memoryless) and adaptive compression (with memory) on the FBE signal $e_{j,6}$ in a normalized scale. It is seen that the use of ACLs emphasizes the onsets and offsets of the FBE signal. In a steady-state portion, the response of adaptive ACLs decays faster than the static polynomial logarithmic compression.

Let us compare the ACDCs with the regression-based velocity and acceleration feature vectors computed from static mMFCCs. For the static mMFCCs, the three main computational stages are: 1) computation of FBEs; 2) static polynomial logarithmic compression; and 3) decorrelation using DCT. As the DCT is a linear operation, we can note that a linear regression operation on mMFCCs across speech frames is equivalent as a linear regression operation on statically compressed FBEs across speech frames. Fig. 5(a) shows the same trajectory of FBE signal for the sixth triangular filter as shown in Fig. 4(b). In Fig. 5(b), we compare the outputs of adaptive compression (ACLs) on the FBE signal $e_{j,6}$ and a linear regression-based velocity operation on the statically (polynomial logarithmic) compressed FBE signal. We note that the velocity operation emphasizes the offsets and onsets of the FBE signal $e_{j,6}$ with a sharp transition. The velocity response decays in a much faster rate than the adaptive ACLs' response and quickly reaches to zero steady state level. For visual clarity, we show the response of velocity operation in Fig. 5(b), but not the response of acceleration operation. The acceleration feature vector is computed through the use of a linear regression method on the velocity feature vector. Hence, the response of acceleration operation is sharper than the response of velocity operation. The sharp decaying responses of velocity and acceleration operations are comparatively faster than the slowly decaying auditory responses. Therefore, we note that the use of ACLs provide additional auditory motivated information that is not captured by the use of the linear regression method.

D. A Note on Computational Complexity

In this subsection, we provide an informal way of comparing computational complexity between ACDCs and the feature vector of [7] that is directly derived from the DAM. For the DAM (see Section III-B and Fig. 1), we note that the input speech signal is decomposed into critical band signals and then ACLs is used to compress each critical band signal. Note that the critical band signals have the same sampling frequency of the input speech signal. On the other hand, for ACDCs, ACLs is used for each FBE signal that has a sampling frequency of speech frame rate. Let us consider typical cases of 16 kHz and 100 Hz (corresponding to 10-ms frame shift) sampling frequencies for input speech signal and FBE signal, respectively. For these typical cases, we can say that the required computation of the feature vector directly derived from the DAM [7] is in the order of 160 times of the required computation of ACDCs.

VI. New Feature Vector

Following standard approach to construct a standard feature vector [25], we append the logarithmic energy of speech frame to the static mMFCCs, and then compute the spectro-temporal feature vectors, such as velocity and acceleration, using a standard regression method across the speech frames. The standard mMFCC feature vector is a combined feature vector that consists of two parts: 1) logarithmic energy of speech frame and static mMFCC feature vector, and 2) their velocity and acceleration feature vectors. The regression method-based velocity and acceleration feature vectors are used to model the underlying dynamic process in the production of speech. On the other hand, the dynamic ACDC feature vector is designed and optimized using the spectro-temporal auditory model DAM. Through ASR experiments (reported in Section VII-B), we observed that the ACDC feature vector is unable to fully replace the existing regression-based velocity and acceleration feature vectors. To construct a standard feature vector by combining the static and dynamic feature vectors, if we use the ACDC feature vector instead of the velocity and acceleration feature vectors, the ASR performance does not reach state-of-the-art performance. Therefore, we propose to use the auditory model-based dynamic ACDC feature vector along-with the regression based dynamic feature vectors.

We use the new ACDC feature vector along-with the standard mMFCC feature vector to construct a new feature vector. The new feature vector is referred to as generalized MFCCs (gMFCCs). Using a 12-dimensional static mMFCCs, we construct a 39-dimensional standard feature vector of mMFCCs. Then we append the 12-dimensional ACDCs to the 39-dimensional mMFCCs to construct the 51-dimensional gMFCCs. Therefore, the vector of 51-dimensional gMFCCs consists of three parts: 1) logarithmic energy of speech frame and 12-dimensional static mMFCC feature vector; 2) their velocity and acceleration feature vectors; and 3) 12-dimensional ACDC feature vector. We note that the new gMFCC feature vector is designed and optimized using both spectral and spectro-temporal auditory models. We also note that the increase in dimensionality of the gMFCC feature vector is reasonable compared to a standard feature vector.

VII. ASR Experiments and Results

Using the HTK software [25], we performed phone and word recognition experiments to show the validity of our distance preserving conjecture followed by performance improvement of new feature vectors. In the next subsections, we first discuss the experimental setups and then show the experimental results.

A. Experimental Setups

We used clean speech training and performed testing for clean speech as well as noisy speech. For extraction of the feature vectors, we used 32-ms frame length and 10-ms frame shift. To achieve better robust recognition performance, we used mean and variance normalization on the full feature vectors on an utterance by utterance basis [29]. For feature vector normalization in robust ASR, the mean and variance normalization method is regarded as a standard method that helps to reduce the statistical mismatch between clean training data and noisy testing data. The feature vector normalization was performed for both training and testing data. The recognition performance is measured in terms of recognition accuracy. In the experiments, we considered that all sound units (phones or words) are equally probable and therefore, no language model was used.

For phone recognition, we used the TIMIT database where the speech is sampled at 16 kHz (wideband speech). The TIMIT database consists of 5300 phonetically balanced clean speech utterances which are partitioned into two sets: a training set consisting of 4620 utterances and a test set consisting of 1680 utterances. HTK training and testing were performed using the training set and the test set of TIMIT, respectively. The TIMIT transcriptions are based on 61 phones. Following convention, the 61 phones were folded onto 39 phones as described in [26]. Using the test set, the testing was performed on 64 145 phone samples. For robust phone recognition experiment, the clean test speech database of TIMIT was corrupted with additive noise. We used the following noise types from the NoiseX-92 [28] database: white, pink, babble, and car ("Volvo") noise. The test speech database was corrupted by adding each noise at 10-dB SNR. The standard configuration of the HTK setup was used where HMMs were trained using three states per phone and 20 Gaussian mixtures per state. We used Gaussian mixtures with diagonal covariance matrices. Unless explicitly specified, we used the mentioned experimental setup as the standard setup for phone recognition experiments.

In the case of word recognition (TIDIGITS), we used the Aurora-2 database where the speech is sampled at 8 kHz (narrowband speech). In the Aurora-2 database, noisy test data sets are available for different noise types at varying SNRs. The noisy test data sets are partitioned into three sub-sets: test set a, test set b, and test set c. For "test set a" and "test set b," the word samples are same, but the noise types are different; the number of words in "test set c" is less than the "test set a" and "test set b." In this paper, we report recognition results for the "test set a" to save space; the performance trend for other two test sets are similar as like the case of "test set a." The "test set a" is further partitioned into four sub-sets: set 1, set 2, set 3, and set 4. The four sub-sets are corrupted by subway, babble, car and exhibition noises respectively at different SNRs. In the Aurora-2 database, the clean speech training data consists of 8440 utterances and the "test set a" consists of 28 028 utterances. Within the "test set a," the four sub-sets consist of 3257, 3308, 3353, and 3241 word samples, respectively. The standard configuration of the HTK setup was used where HMMs were trained using 16 states per word and three Gaussian mixtures per state (diagonal covariance matrices).

B. Experimental Results

We first show the phone recognition results using 12-dimensional mMFCC feature vector to validate the conjecture of local geometry preservation between the feature domain and the auditory domain. For wideband speech (16-kHz sampling frequency), we found the optimal parameters of mMFCCs that minimizes the objective measure ξ of (1). In Section IV-A, we evaluated the optimal warping factor $\alpha = 900$ and the optimal values of the polynomial coefficients as $b_1 = 0.1$ and $b_2 = 0.9$. For $\alpha = 900$ and considering $b_1 + b_2 = 1$, any value of b_1 except 0.1 should worsen the objective measure ξ and lead to poorer recognition performance if the conjecture holds true. Therefore, to validate the conjecture, we carried out phone recognition experiments where the parameter b_1 is varied keeping the parameter α fixed. Using the standard experimental setup, the phone recognition results are shown in Table I for varying $b_1 \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1\}$ where $\alpha = 900$. The set of b_1 was chosen arbitrarily. For robust recognition, we experimented with white noise corrupted test dataset. The white noise was chosen as it affects the power spectrum throughout the full frequency range. From Table I, we note that $b_1 = 0.1$ provides the best performance. Note the poorest performances at $b_1 = 1$, specially in the case of white noise case. A natural question is how statistically significant is the optimal value of the parameter $b_1 = 0.1$ with respect to the other values. For computing the significance levels, the full test set was divided into eight sub sets and then the recognition results were evaluated for all the subsets at different values of b_1. Using the set of recognition scores, we computed the significance levels through a standard T-test for which the null hypothesis is rejected in favor of the optimal choice of the $b_1 = 0.1$ against the choice of other values and the significance levels are shown in Table II. We do not report the recognition results for all the eight subsets at different values of b_1 for brevity of the paper. For the mMFCCs, the experimental evaluation verifies the optimal choice of $b_1 = 0.1$ at $\alpha = 900$ that was obtained using the perturbation-based approach of preserving the local geometries. Thus, the information that our method extracts

TABLE I
USING 12-DIMENSIONAL STATIC mMFCCs: PHONE RECOGNITION ACCURACY (IN %) AGAINST CHOICES OF b_1 FOR THE FIXED $\alpha = 900$

Test Condition	b_1					
	0.01	0.05	0.1	0.2	0.5	1
Clean speech	53.75	53.64	54.20	53.32	53.02	52.02
White noise (SNR = 10 dB)	35.50	36.75	36.79	35.63	32.90	31.17

TABLE II
SIGNIFICANCE LEVELS FOR THE OPTIMAL CHOICE OF $b_1 = 0.1$ WITH RESPECT TO THE OTHER b_1 VALUES

b_1				
to 0.01	to 0.05	to 0.2	to 0.5	to 1
Clean speech				
7×10^{-3}	3×10^{-3}	5×10^{-8}	3×10^{-6}	10^{-6}
White noise at SNR=10 dB				
6×10^{-6}	0.35	3×10^{-7}	2×10^{-9}	9×10^{-11}

TABLE III
PHONE RECOGNITION ACCURACY (IN %) OF STATIC 12-DIMENSIONAL MFCC AND mMFCC FEATURE VECTORS FOR VARYING NUMBER OF GAUSSIAN MIXTURES/STATE

Feature	Number of Gaussian mixtures/state						
	1	2	4	8	12	16	20
	Clean speech test condition						
MFCC	44.58	46.68	48.12	50.25	51.24	52.07	52.33
mMFCC	46.27	48.55	50.34	52.24	53.27	53.87	54.20
	White noise test condition; SNR = 10 dB						
MFCC	31.54	31.38	31.22	31.01	30.90	31.15	31.22
mMFCC	35.89	35.97	36.31	36.56	36.71	36.73	36.79

TABLE IV
PHONE RECOGNITION ACCURACY (IN %) OF DYNAMIC FEATURE VECTORS FOR CLEAN SPEECH TESTING

Feature Vectors (dimension)	
Velocity of mMFCCs (12)	47.84
Acceleration of mMFCCs (12)	37.49
Velocity and acceleration of mMFCCs (24)	56.93
ACDCs (12)	47.26
ACDCs along-with velocity and acceleration of mMFCCs (36)	60.67

TABLE V
PHONE RECOGNITION ACCURACY (IN %) OF DYNAMIC FEATURE VECTORS FOR VARYING NUMBER OF GAUSSIAN MIXTURES/STATE

Feature (dimension)	Number of Gaussian mixtures/state						
	1	2	4	8	12	16	20
	Clean speech test condition						
RDFV (24)	35.55	44.44	49.32	53.47	55.17	56.47	56.93
CDFV (36)	44.56	50.71	54.17	57.33	58.94	59.86	60.67
	White noise test condition; SNR = 10 dB						
RDFV (24)	29.77	33.45	35.70	37.76	38.53	38.92	38.97
CDFV (36)	36.28	37.85	39.67	41.74	42.85	43.54	43.80

from the auditory model and unlabeled data results in optimal practical recognition performance.

Next we considered the issue of how the recognition performance of 12-dimensional static mMFCC vector fares with respect to the 12-dimensional static MFCC vector at varying experimental setup. This experiment helps us to observe the effect of an experimental setup on the recognition performance of static feature vectors. The static 12-dimensional MFCC feature vector was extracted using the same setup as that used to extract the 12-dimensional mMFCC feature vector (for MFCC, $\alpha = 700$ and $b_1 = 1$). Keeping the number of HMM states/phone fixed, we considered to compare between 12-dimensional MFCCs and mMFCCs for phone recognition where the number of Gaussian mixtures/state varies. Note that the use of more number of mixtures/state requires more computational resource. Table III shows the recognition results for varying number of mixtures where the number of states/phone is kept as three. We observe that the static mMFCCs are consistently better than the static MFCCs for any number of mixtures/state. In the case of the white noise test condition, the mMFCCs are significantly better than the MFCCs.

Further we compared between the dynamic feature vectors: the auditory model based ACDCs and the linear regression-based velocity and acceleration feature vectors, and their combinations. The linear regression based feature vectors are computed from the static mMFCCs. Using the standard setup, Table IV shows the comparative study between several dynamic feature vectors and their combinations for the task of clean speech phone recognition. We note that the 12-dimensional ACDC feature vector is unable to compete with the linear regression based 24-dimensional dynamic feature vector that consists of the velocity and acceleration feature vectors. The 12-dimensional ACDC feature vector provides similar performance to the 12-dimensional velocity feature. To achieve better performance, we combined the 12-dimensional ACDC feature vector and the regression based 24-dimensional dynamic feature vector to create a new 36-dimensional dynamic feature vector. From Table IV, we note that the new 36-dimensional dynamic feature vector provides noticeable improvement in the recognition performance compared to the regression based 24-dimensional dynamic feature vector. Let us refer to the linear regression based 24-dimensional dynamic feature vector as the regression-based dynamic feature vector (RDFV) and the new 36-dimensional dynamic feature vector as the combined dynamic feature (CDFV). We considered the issue of how the recognition performance of 24-dimensional RDFV fares with respect to the 36-dimensional CDFV at varying experimental setup. Using three states/phone, we compared between RDFV and CDFV for phone recognition where the number of Gaussian mixtures/state was varied and the recognition results are shown in Table V. For clean speech phone recognition, the 36-dimensional CDFV at eight mixtures/state provides similar performance to the 24-dimensional RDFV at 20 mixtures/state. The performance improvement of CDFV can be attributed to the use of auditory motivated dynamic ACDCs. Therefore, to construct the gMFCC as a standard feature vector (see Section VI), we proposed to use the dynamic ACDCs along-with the standard set of static and regression based dynamic feature vectors.

Finally, we considered the comparison of standard feature vectors using phone and word recognition experiments. Using the standard approach, 39-dimensional feature vectors were evaluated for the cases of MFCC and mMFCC feature vectors. To the static feature vector, we appended the log energy of the corresponding speech frame and the velocity and acceleration

TABLE VI
PHONE RECOGNITION ACCURACY (IN %) FOR STANDARD FEATURE VECTORS

Feature	Clean	Noise Types; SNR=10 dB				Average
(dimension)		White	Pink	Babble	Volvo	Performance
MFCC (39)	68.11	37.03	40.51	46.25	59.71	50.32
RASTA-PLP (39)	67.00	40.39	41.41	47.38	62.01	51.63
mMFCC (39)	68.34	43.65	46.67	48.94	61.90	53.90
gMFCC (51)	68.54	43.86	47.34	48.92	63.31	54.39

TABLE VII
SIGNIFICANCE LEVELS FOR PHONE RECOGNITION RESULTS AS SHOWN IN TABLE VI: gMFCC WITH RESPECT TO MFCC, RASTA-PLP AND mMFCC

gMFCC to MFCC	gMFCC to RASTA-PLP	gMFCC to mMFCC
1.1×10^{-2}	1.3×10^{-2}	6.5×10^{-2}

feature vectors.[7] In case of a 51-dimensional gMFCC feature vector, a 12-dimensional ACDC vector is appended to the corresponding 39-dimensional mMFCC feature vector (see Section VI). We also show the performance of 39-dimensional relative spectral perceptual linear prediction (RASTA-PLP) coefficients [31], [33]. Following [34], for the RASTA-PLP feature vector, a 13-dimensional static feature vector was extracted, and the velocity and acceleration feature vectors were computed followed by appending them to the static feature vector to construct the 39-dimensional feature vector. Therefore, we compared between 39-dimensional MFCC, mMFCC, RASTA-PLP feature vectors and 51-dimensional gMFCC feature vector. For phone recognition, the performance of the feature vectors are shown in Table VI using the standard experimental setup. The RASTA-PLP is found to better than MFCC feature vector for all noise types. For white noise, the RASTA-PLP shows more than 3% improvement over MFCC feature vector. We note that mMFCC feature vector performs better than MFCC and RASTA-PLP feature vectors for all noise types without any penalty of increasing dimensionality. For white and pink noise cases, we note that the mMFCC feature vector provides substantial improvement over MFCC and RASTA-PLP feature vectors. The use of the mMFCC feature vector provides absolute average performance improvement of $(53.90 - 50.32)\% = 3.58\%$ over the MFCC feature vector. The gMFCC feature vector provides slight improvement over mMFCC feature vector. Overall, using the gMFCC feature vector, we achieve an absolute average performance improvement of 4.07% over the standard MFCC feature vector and 2.76% over the standard RASTA-PLP feature vector. Significance levels associated with Table VI are shown in Table VII where performance of the gMFCC feature vector is used as the reference score.

Next, we performed word recognition using the Aurora-2 database and show the robust recognition results in Table VIII at varying SNR conditions (using the standard experimental setup). RASTA-PLP is found to be better than MFCC feature vector. We note that the mMFCC feature vector performs better than the MFCC feature vector and similarly to the RASTA-PLP feature vector. The gMFCC feature vector provides further improvement on the mMFCC feature vector. For SNR $=$ 10 dB, the mMFCC feature vector provides an absolute average performance improvement of 2.80% over the MFCC feature vector and the gMFCC feature vector provides further absolute average performance improvement of 3.52% over the mMFCC feature vector. Overall, using the gMFCC feature vector, we achieve an absolute average performance improvement of 6.32% over the standard MFCC feature vector for SNR $=$ 10 dB, but at the expense of a moderate increase in dimensionality. Comparing with RASTA-PLP at SNR $=$ 10 dB, the gMFCC feature vector provides 2.89% absolute average performance improvement. We note that the mMFCC feature vector provides noticeable improvement over the standard MFCC feature vector without any penalty of increasing dimensionality. Significance levels associated with Table VIII are shown in Table IX, where performance of the gMFCC feature vector is used as the reference score.

TABLE VIII
WORD (TIDIGITS) RECOGNITION ACCURACY (IN %) FOR STANDARD FEATURE VECTORS

Feature	Test Set a				Average
(dimension)	set 1	set 2	set 3	set 4	Performance
	Clean speech testing				
MFCC (39)	99.05	98.91	98.99	99.11	99.01
RASTA-PLP (39)	99.23	99.15	99.08	99.32	99.19
mMFCC (39)	99.26	98.97	99.05	99.29	99.14
gMFCC (51)	99.32	99.00	99.28	99.38	99.24
	Noisy speech testing				
	Noise types				
	Subway	Babble	Car	Exhibition	
	SNR = 20 dB				
MFCC (39)	95.46	96.67	96.12	94.88	95.78
RASTA-PLP (39)	96.96	97.88	97.29	96.42	97.13
mMFCC (39)	96.59	97.49	97.08	96.42	96.89
gMFCC (51)	97.73	98.19	97.91	97.72	97.88
	SNR = 10 dB				
MFCC(39)	85.05	86.49	83.39	81.70	84.15
RASTA-PLP (39)	88.58	89.03	87.29	85.44	87.58
mMFCC (39)	87.07	88.51	87.00	85.25	86.95
gMFCC (51)	90.91	90.99	90.69	89.29	90.47

TABLE IX
SIGNIFICANCE LEVELS FOR WORD RECOGNITION RESULTS AS SHOWN IN TABLE VIII: gMFCC WITH RESPECT TO MFCC, RASTA-PLP AND mMFCC

gMFCC to MFCC	gMFCC to RASTA-PLP	gMFCC to mMFCC
1.1×10^{-3}	3.1×10^{-3}	1.7×10^{-3}

VIII. CONCLUSION

Our development of new feature vectors shows that the judicious use of sophisticated auditory models can lead to simple feature vectors that provide improved speech recognition performance for any environmental condition. We note that both spectral and spectro-temporal auditory properties are helpful to design improved feature vectors. The success of our perceptual-distance preserving measure in optimizing feature vectors suggests that the auditory system provides as output a signal representation that is "efficient" for speech recognition in the sense of providing relevant information.

[7] The 39-dimensional MFCC or mMFCC feature vector refers to the corresponding feature vector where we appended the log energy of a speech frame to the corresponding 12-dimensional static feature vector to make a 13-dimensional feature vector. Then we computed the velocity and acceleration feature vectors from the 13-dimensional feature vector and appended them to make a 39-dimensional feature vector.

In the development of auditory motivated new dynamic feature vector, the use of a memory circuit (ACLs) of a spectro-temporal auditory model is shown to provide complimentary auditory information across the speech frames in the sense of memory and hence, assisting in the ASR task.

Acknowledgment

The authors would like to thank V. Surayanarayana. K of the Indian Institute of Science for his help to set up HTK and the Medical Physics group at the Carl von Ossietzky–Universitat Oldenberg for code fragments implementing the adaptive compression loops. The authors would also like to thank the anonymous reviewers for their constructive suggestions.

References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[2] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.

[3] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, no. 6, pp. 2623–2629, Jun. 1989.

[4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[5] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 451–464, Sep. 1997.

[6] D. S. Kim, S. Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[7] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 2040–2050, Oct. 1999.

[8] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 1, pp. 43–49, Jan. 2006.

[9] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proc. ICASSP*, 2009, pp. 4625–4628.

[10] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *Proc. ICASSP*, 2010, pp. 4514–4517.

[11] S. Seneff, "A joint synchrony/mean-rate model of auditory processing," *J. Phonet.*, vol. 85, no. 1, pp. 55–76, Jan. 1988.

[12] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, vol. 79, no. 3, pp. 702–711, Mar. 1988.

[13] J. M. Kates, "Two-tone suppression in a cochlear model," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 396–406, Sep. 1995.

[14] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, Jun. 1996.

[15] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3623–3631, Jun. 1996.

[16] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modelling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Amer.*, vol. 102, no. 5, pp. 2892–2905, Nov. 1997.

[17] A. J. Oxenham, "Forward masking: Adaptation or integration?," *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 732–741, Feb. 2001.

[18] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A Perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1292–1304, 2005.

[19] B. K.-W. Mak, Y.-C. Tam, and P. Q. Li, "Discriminative auditory-based features for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 12, no. 1, pp. 27–36, Jan. 2004.

[20] W. R. Gardner and B. D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 367–381, Sep. 1995.

[21] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 310–319, Jan. 2007.

[22] C. Koniaris, M. Kuropatwinski, and W. B. Kleijn, "Auditory-model based robust feature selection for speech recognition," *J. Acoust. Soc. Amer. Exp. Lett.*, vol. 127, no. 2, pp. EL73–EL79, Feb. 2010.

[23] C. Koniaris, S. Chatterjee, and W. B. Kleijn, "Selecting static and dynamic features using an advanced auditory model for speech recognition," in *Proc. ICASSP*, 2010, pp. 4342–4345.

[24] S. Chatterjee, C. Koniaris, and W. B. Kleijn, "Auditory model based optimization of MFCCs improves automatic speech recognition performance," in *Proc. Interspeech*, Sep. 2009, pp. 2987–2990.

[25] S. Young *et al., The HTK Book (for HTK Version 3.4)*. Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2006.

[26] 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.

[27] S. Ganapathy, S. Thomas, and H. Hermansky, "Modulation frequency features for phone recognition in noisy speech," *J. Acoust. Soc. Amer. Exp. Lett.*, vol. 125, no. 1, pp. EL8–EL12, Jan. 2009.

[28] NoiseX-92, Rice Univ. DSP group, 2009, accessed Feb. 2009.

[29] J. Droppo and A. Acero, "Environmental robustness," in *Handbook of Speech Processing*. New York: Springer, 2007, pp. 658–659.

[30] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, no. 3, pp. 153–181, 1957.

[31] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[32] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, Sep. 2009, pp. 28–31.

[33] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Apr. 1994.

[34] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai-Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. ICASSP*, 2008, pp. 4449–4452.

Saikat Chatterjee (M'09) received the Ph.D. degree from the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India.

He is a Researcher in the Communication Theory Lab, KTH-Royal Institute of Technology, Stockholm, Sweden, where he also pursued one year post-doctoral study. He was also with the Sound and Image Processing Lab at the same institution as a Post-Doctoral Fellow for one year.

Dr. Chatterjee was a coauthor of the paper that won the Best Student Paper Award at ICASSP 2010. His current research interests are source coding, speech and audio processing, estimation and detection, sparse signal processing, compressive sensing, and wireless communications.

W. Bastiaan Kleijn (M'88–SM'97–F'99) received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, the M.S. degree in physics and the Ph.D. degree in soil science, both from the University of California, Riverside, and the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands.

He has been a Professor of Electronic Engineering at Victoria University of Wellington, Wellington, New Zealand since 2010. He is also a Professor at the School of Electrical Engineering, KTH (Royal Institute of Technology), Stockholm, Sweden, where he was until recently Head of the Sound and Image Processing Laboratory. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996. He was a founder of Global IP Solutions, which was acquired by Google in 2010.

Prof. Kleijn is on the Editorial Board of *Signal Processing* and has been on the Boards of the IEEE Transactions on Speech and Audio Processing, IEEE Signal Processing Letters, IEEE Signal Processing Magazine, and the *EURASIP Journal of Applied Signal Processing*. He has been a member of several IEEE technical committees, and a Technical Chair of EUSIPCO 2010, ICASSP'99, the 1997 and 1999 IEEE Speech Coding Workshops, and a General Chair of the 1999 IEEE Signal Processing for Multimedia Workshop.