

## Chapter

# Generalized Spectral-Temporal Features for Representing Speech Information

*Stephen A. Zahorian, Xiaoyu Liu and Roozbeh Sadeghian*

## Abstract

Based on extensive prior studies of speech science focused on the spectral-temporal properties of human speech perception, as well as a wide range of spectral-temporal speech features already in use, and motivated by the time-frequency resolution properties of human hearing, this chapter proposes and evaluates one general class of spectral-temporal features. These features, intended primarily for use in Automatic Speech Recognition (ASR) front ends, allow different realizations of general time-frequency concepts to be easily implemented and tuned through a set of frequency and time-warping functions. The methods presented are flexible enough to allow evaluation of the relative importance of the spectral and temporal features and to explore the trade-off between time and frequency resolution. Extensive ASR experiments were conducted to evaluate various spectral-temporal properties using this unified framework.

**Keywords:** time-frequency, features, automatic speech recognition, basis vectors, front end

## 1. Introduction

As mentioned elsewhere [1], good features for automatic speech recognition include relevance, compactness, completeness, and robustness. That is, speech features should be closely related to speech production and understanding, should be small in number, represent as much speech information as possible, and should be little changed in the presence of noise or varying external conditions.

As these elements suggest, both productive and receptive aspects of speech science form the foundation for signal processing to extract speech features. Although receptive aspects of speech science are most directly relevant to speech features for ASR, speech production models for vocal tract configurations are also a plausible starting point for guiding speech feature extraction. In terms of speech production, ever since the classic Peterson and Barney vowel study [2], by far the most widely used acoustic features for characterizing vocal tract shape are formants. For speech signal

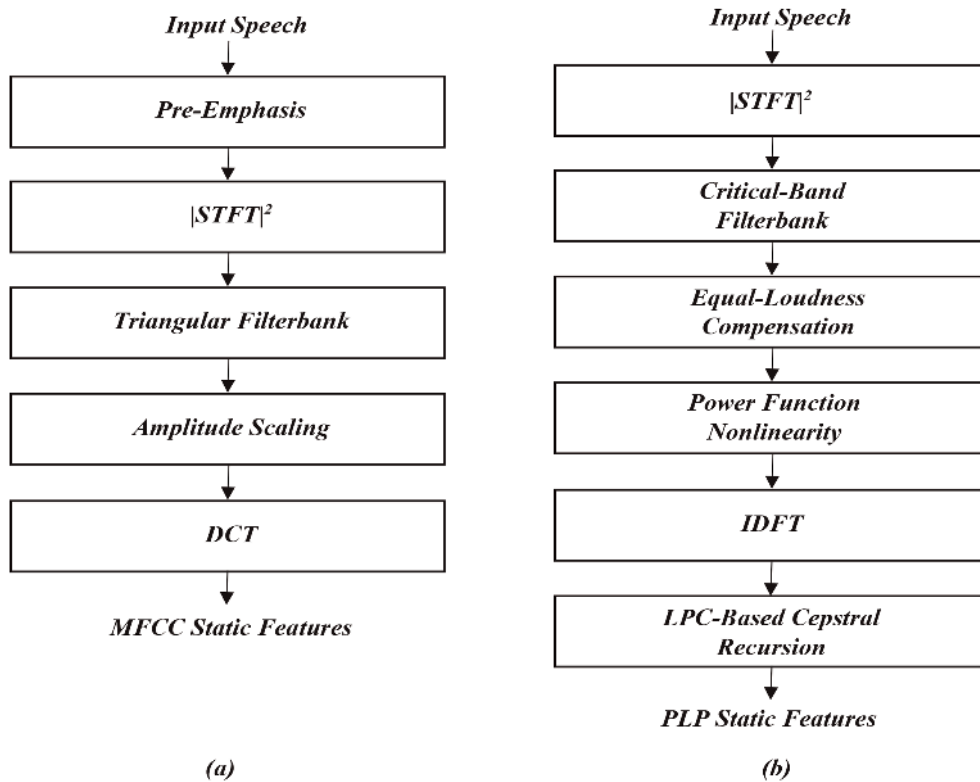
processing applications, formant information is generally obtained by first modeling the vocal tract using an all-pole system, such as in the Perceptual Linear Predictive (PLP) front end [3]. The motivating idea is that nearly any transfer function can be approximated by a high-order all-pole model. Due to lack of automatic methods to reliably estimate formants [4], and also because formants cannot discriminate between speech sounds for which the main differences are unrelated to formants (such as fricatives) [5, 6], formants are rarely used as features for ASR. For ASR the all-pole approximation to the vocal tract is more typically replaced with cepstral features [7], which encode the global spectral envelope shape without any emphasis given to spectral peaks.

There are many complex issues raised in the speech science literature about receptive aspects of human speech that could be potentially taken into account for extracting speech features for use in ASR. However, the only effects taken into account for the features presented in this chapter are the primary considerations of frequency and temporal resolution.

Auditory processing research related to the cochlea's frequency selectivity provides the fundamental theory for auditory filterbanks, which are often used as a signal processing step to compute features for ASR. Many canonical studies, such as [8–10], have pointed out that humans discern low frequency components in a complex sound with much higher resolution than is the case for high frequencies. Hence, in speech front ends, to mimic this property, the physical frequency range is mapped to a perceptual scale, typically using bandpass filtering with 25–60 overlapping bands, each corresponding to approximately equal length regions along the cochlear membrane. The bandwidths are designed to match the frequency resolution at each center frequency. Various perceptual scales have been developed, such as the Mel scale [9], Bark scale [10, 11], and Equivalent Rectangular Bandwidth (ERB) scale [12].

Commonly used filterbanks include triangular filters [13] based on the Mel scale, trapezoidal filters [3] based on the Bark scale, and gammatone filters [14, 15] based on the ERB scale. The output power of each filterbank channel is computed as a weighted sum of the magnitude-squared Short Time Fourier Transform (STFT), weighted by the channel frequency response, and then amplitude scaled to approximate perceptual loudness, which is linearly proportional to the neuron firing rate of the auditory nerves [16]. The amplitude-scaled outputs are usually combined with a cosine transform to form cepstral features such as the widely used MFCC features [13]. Another front end for computing speech features is PLP [3]. In PLP an equal-loudness compensation is also modeled to account for the non-equal amplitude sensitivity of human hearing at different frequencies [17]. Motivated by the importance of formants, linear prediction coefficients are computed from the Bark domain spectrum using Durbin's recursive method [18] and then converted to cepstral features.

**Figure 1** depicts static feature extraction for the MFCC and PLP front ends. Note that the expression static features refers to features computed from a single very short segment of speech (on the order of 20 ms duration), called a frame. These features are computed for each frame with frames typically spaced apart by approximately 10 ms, thus also overlapped by 10 ms. This gap between adjacent frames is the frame spacing. Static features based on perceptual frequency scales do not do not explicitly encode spectral trajectories over time. In [19–22] approximations of time “derivatives” of the static features are computed and appended to static features to reduce ASR error rate considerably (empirically on the order of 20%). These time derivatives are called



**Figure 1.**  
 Comparison of the MFCC (a) and PLP (b) structure.

dynamic features and are also often referred to as delta and acceleration (second order differential) terms. Mathematically, the delta terms are computed as:

$$\Delta_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (1)$$

Where  $\Delta_t$  is the differential at time  $t$  estimated from small adjacent groups of static features (cepstrums)  $c_{t-\theta}$  to  $c_{t+\theta}$  with  $2\Theta + 1$  being the total number of surrounding frames. In the remainder of this chapter, groups of frames used to compute dynamic features from static features are referred to as blocks. More detailed discussion of frames and blocks, specifically related to the spectral-temporal features presented in this chapter, is given in Section 2.

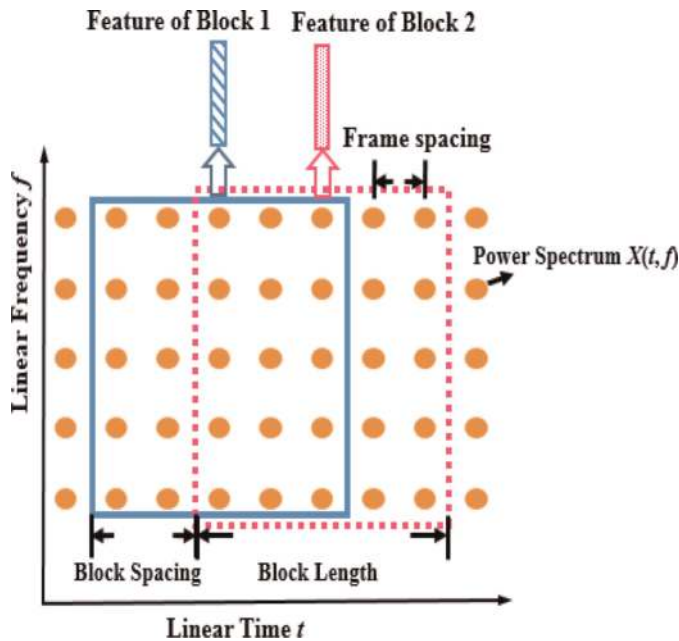
Note that although the time derivatives are estimated from a short block of features, they essentially characterize the spectral trajectory at each single time instant, and thus are unable to account for the non-uniform time resolution of the human auditory system observed over a long duration of time. Spectral-temporal modulation features are much more effective than the delta method in addressing the issue of non-uniform time-frequency resolution and efficiently sampling the short-time spectrum. In 1994 Drullman et al. [23] found that the most important spectral trajectory information over time for speech perception is in the range of 1–16 Hz “modulation” frequencies. Guided by this finding, in order to exploit the information in the modulation frequencies, relatively long time blocks of each spectral band are analyzed. Over many years, various modulation features have been investigated.

Athineos et al. [24] used the dual of time-domain linear prediction to frequency-domain model the poles of the temporal envelope in each sub-band. Valente and Hermansky [25] developed an approach combining independent classifier outputs and modulation frequency channels. Gabor-filter-based approaches for extracting localized directional features also show promise [26, 27]. However, the large number of parameters, which allow Gabor filters to be aligned in many different directions, presents the added difficulty of determining these directions in an effective way for use in ASR.

Based on this prior extensive groundwork, this chapter presents a generalized spectral-temporal feature extraction front end for representing speech information. This feature set encompasses a wide range of time-frequency representation options focusing on two important properties of human hearing—frequency and time resolution. Rather than presenting one specific type of front end, a unified framework is presented such that various realizations of the general time-frequency concepts can easily be implemented and tuned. Based on a set of frequency-warping and time-warping functions, this front end is flexible enough to allow straightforward evaluation of the trade-off between frequency and time resolution at the acoustic feature level.

## 2. Method

The spectral-temporal features presented in this chapter are weighted sums of short-time spectral magnitudes, using overlapping frame-based processing. **Figure 2** illustrates the division of the short-time spectrum. The horizontal and vertical axes represent physical time (in seconds) and physical frequency (in Hz). A time-frequency representation (TFR) of the speech, denoted by  $X(t, f)$ , is obtained by computing the magnitude-squared STFT of each frame. In **Figure 2**, the dots in each



**Figure 2.** A high level illustration of the proposed front end and definitions of related terminologies.

column represent the power spectrum of a frame, and the gap between adjacent columns denotes the frame spacing. Note that unlike the MFCC or PLP front ends, for which each feature vector is the concatenation of the spectral (static) feature and the spectral trajectory (dynamic feature) components, and the spectral trajectory is characterized by the time derivatives of the static terms at each sample instant on a frame-by-frame basis, in the method presented in this chapter, the front end computes a set of spectral-temporal features for a long block of spectral values centered at each sample instant, and one feature vector is extracted for each block. As will be seen in the derivations, this spectral-temporal feature vector for each block integrates both the spectral and temporal aspects of the speech signal within the block by a weighted sum of  $X(t, f)$  based on a set of two-dimensional spectral-temporal basis vectors. Thus, in the proposed front end, there are no individual static components in the final features since they are fused in the output features. Also, the use of long segments to compute features, using short highly overlapped frames, non-uniform time resolution can be incorporated in spectral trajectories.

Two basic concepts are also illustrated in **Figure 2**, which are used and referred to in the remainder of this chapter—block length and block spacing. Block length is defined as the time duration (physical time) of a block of short-time frames. Block length is measured in milliseconds and is equal to the frame spacing multiplied by the number of frames in the block. The spacing between two adjacent blocks is defined as block spacing, which is the product of the frame spacing and the number of frames that separate the two blocks. Since features are extracted on a block basis, the block spacing is also the feature spacing. At the beginning and ending of each speech utterance, zero padding is used to allow the first and last blocks to be centered at the first and last frames respectively. As opposed to MFCC or PLP processing, in which the feature spacing is identical to the frame spacing, in our work the feature spacing is typically considerably larger than the frame spacing. With these high level concepts, a detailed illustration of the feature extraction process is presented in the remainder of this section.

The time-frequency plane obtained by STFT has uniform frequency and time resolution determined by the analysis window shape and width [28]. This representation does not take into account the non-uniform perceptual frequency scale of the peripheral auditory system. For convenience and clarity of explanation, a framework is established with  $t'$  and  $f'$  as normalized perceptual time and frequency scales, whose desirable properties are next described in detail. Then a set of features,  $Feat(i, j)$  for the time block centered at time instant  $t$ , can be expressed as:

$$Feat(i, j) = \int_{t'=-1/2}^{1/2} \int_{f'=0}^1 a(X'(t', f')) \cdot BV_{i,j}(t', f') df' dt'. \quad (2)$$

In Eq. (2) the feature computation is performed using perceptual scales, where  $X'(t', f')$  is the power spectrum of a time-frequency block in this domain for which the frequency  $f'$  is mapped to the range of  $\{0, 1\}$  by subtracting an offset and dividing by a scaling factor. Similarly, perceptual time  $t'$  is converted to the range of  $\{-1/2, 1/2\}$  with  $t' = 0$  the center of the time block. The function  $a(\cdot)$  nonlinearly maps the power spectrum to a perceptual-loudness scale, most often using a logarithmic scaling or a power-law nonlinearity [29]. Finally, the amplitude-scaled power spectrum is weighted by a set of two-dimensional basis vectors  $BV_{i,j}$  in the perceptual domain  $(t', f')$ . The number of features extracted from a time-frequency block depends on the number of basis vectors used.



It should be emphasized, that for clarity of explanation, integrals as well as continuous time and frequency variables are used in Eq. (2) in all of the following equations. In actual implementations, both time and frequency variables are discrete, as shown in **Figure 2**, and integrations are computed as sums. Also, although the feature extraction is effectively performed in the perceptual time-frequency domain  $(t', f')$ , the actual computations use the linear time-frequency plane. The mapping between linear and perceptual domains for time and frequency are established by nonlinear time and frequency-warping functions and incorporated by changes in underlying basis vectors as explained below.

In this work, a set of two-dimensional cosine basis vectors for  $BV_{i,j}(t', f')$  is used to compactly encode the spectral envelope as well as the spectral trajectory. The theoretical work of Rao and Yip [30] gives reasons why the cosine transform is particularly appropriate for data compression and feature de-correlation, based on similarity to the data-driven Karhunen-Loeve Transform. For similar reasons, the MFCC features also use a one-dimensional cosine transform as a processing step. The popular JPEG standard for image compression also uses two-dimensional cosine transforms.

Continuing with the specifics of the method presented in this chapter, the 2-D cosine basis vectors operating in the perceptual space are defined as:

$$BV_{i,j}(t', f') = \cos(\pi i f') \cdot \cos(\pi j t'), \quad (3)$$

$$0 \leq i \leq N - 1, 0 \leq j \leq M - 1.$$

Eq. (3) shows that each 2-D basis vector is the product of two individual basis vectors, one over frequency  $f'$ , and one over time  $t'$ . The numbers of basis vectors over frequency and time are specified by  $N$  and  $M$  respectively. The total number of features for each block is given by  $N \times M$ . As is discussed in detail in Section 3, a larger  $N$  or  $M$  provides a more detailed representation of the spectral envelope over frequency or the spectral trajectory over time respectively. Empirical data indicates a total of 75 features for each block ( $N = 15, M = 5$ ) results in high ASR accuracy. Eqs. (4) through (9), and associated figures, show that the nonlinear mapping from  $f$  to  $f'$  and  $t$  to  $t'$ , together with their differentials  $df'$  and  $dt'$ , approximate the frequency and time resolution of human hearing. Next is shown how the nonlinear mappings are mathematically incorporated into the feature calculations. Frequency warping, specifies the relation between perceptual frequency  $f'$  and physical frequency  $f$ :

$$f' = g(f), 0 \leq f \leq 1 \quad (4)$$

The physical frequency range has also been normalized to  $\{0,1\}$ <sup>1</sup>. Thus, the  $df'$  term in Eq. (2) is equivalent to:

$$df' = \frac{dg}{df} df \quad (5)$$

---

<sup>1</sup> For convenience, the normalized frequency range  $\{0,1\}$  of  $f$  corresponds to the physical range  $\{0, Fs/2\}$  where  $Fs/2$  is the Nyquist frequency. The normalized perceptual frequency  $f'$  over  $\{0,1\}$  also represents the range of 0 to  $Fs/2$ . With minor changes, this normalized range can be reduced to a shorter frequency range of physical frequencies.

As per the discussion in Section I, one reasonable choice for the form of the frequency warping  $g(f)$  is a Mel-shape warping defined as:

$$g(f) = C \cdot \log_{10} \left( 1 + \frac{f}{k} \right) \quad (6)$$

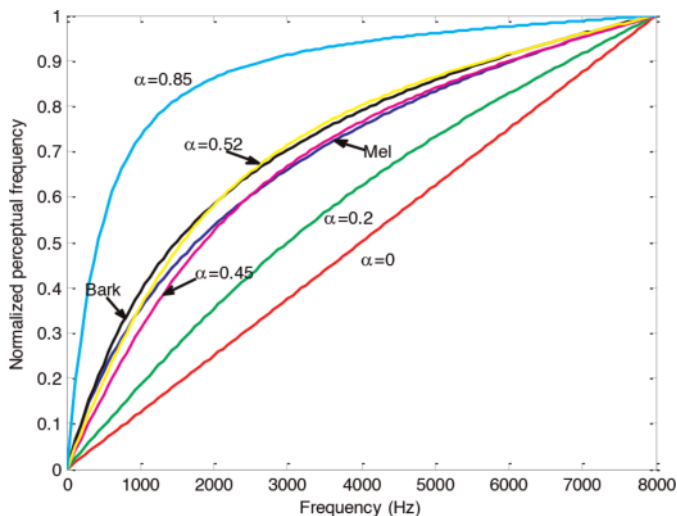
where  $k$  is an adjustable warping factor between 0 and 1 that controls the degree of the warping, and the constant  $C$  is chosen to ensure that  $f = 1$  is mapped to  $f' = 1$ . If  $k = 0.0875$  and  $C = 0.9137$ , for the frequency range of 0 to 8000 Hz, this warping is the normalized version of the most widely used “standard” Mel warping proposed by O’Shaughnessy [31]. Another option, using Smith and Abel’s work [32], is to use a bilinear warping to approximate the Bark scale:

The warping factor  $\alpha$  ranges from 0 to 1. In **Figure 3**, five bilinear warpings, for various  $\alpha$  values, are shown. Additionally, Mel warping using O’Shaughnessy’s equation in [31] and Bark warping as per Wang et al. [33] are plotted in the figure. The figure clearly shows that bilinear warping can be adjusted to closely approximate both Mel and Bark warping. From Eq. (5), frequency resolution is continuously varied, to match auditory properties, rather than using a quantized version with a filterbank, such as in the MFCC, PLP or gammatone front ends, [3, 13, 14]. In the filterbank methods, perceptually indistinguishable frequency components are modeled by the filter bandwidths. Thus, a filterbank is effectively a quantizer which separates the perceptual frequency scale into a finite number of equal intervals. In the proposed approach, the perceptual scale is continuous. The frequency selectivity is modeled by the derivative term  $dg(f)/df$ .

Next, the relation between perceptual time  $t'$  and linear time  $t$  is modeled with nonlinear (warping) function,  $h$ , but with a normalized range of  $t \in \{-1/2, 1/2\}$ :

$$t' = h(t, f); \quad -\frac{1}{2} \leq t \leq \frac{1}{2}, \quad 0 \leq f \leq 1. \quad (7)$$

Time  $t'$  can be considered a perceptual time scale that defines a “pseudo” time instant at which an acoustic event occurring at physical time  $t$  is perceived by the



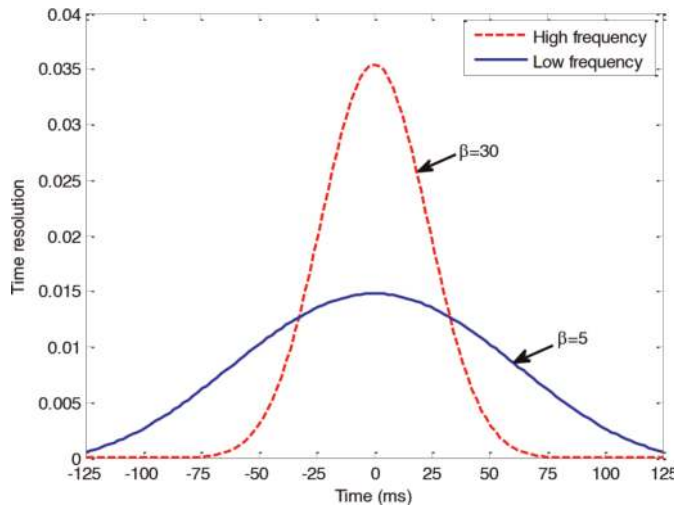
**Figure 3.** Bilinear warping with different warping factors—Mel and Bark warping shown for comparison.

auditory system. Mathematically, perceptual time is given in terms of its derivative with respect to  $t$ :

$$dt' = \frac{dh(t,f)}{dt} dt \quad (8)$$

This time resolution term indicates how far apart two events are perceived when separated by unit time on the physical scale. A large derivative implies that two acoustic events are clearly perceptually distinguishable whereas a small value corresponds to a time boundary between events that is not well resolved. When characterizing the temporal trajectory of acoustic events, it's reasonable to assume that perceptual time resolution should be higher near the center of the event than at far away times. That is, to identify the content of a segment with the help of its left and right segments, it is plausible that close segments are more relevant than far-away segments. Hence, temporal changes of the spectrum envelope should be more clearly resolved at the center of an event than far-away less helpful parts. Therefore, the shape for  $dh/dt$  was chosen to be approximately Gaussian. More specifically,  $dh/dt$  is a Kaiser window, with one parameter,  $\beta$ , the time-warping factor, that conveniently controls the “sharpness” of time warping.

Note that in Eqs. (8), (9) the sharpness of the time resolution term  $dh/dt$  could be frequency dependent as well. Specifically, the term  $dh/dt$  can be made more “peaky” at high frequencies than at low frequencies, controlled by different warping factor values in the Kaiser window<sup>2</sup>, as illustrated in **Figure 4**. This allows an exploration of the trade-off between auditory frequency and time resolution. The psychoacoustic masking experiments [34] show that the very narrow auditory filter bandwidths at low frequencies produces high frequency resolution, but also prolongs the “ring” time at the onset and offset transients for short signals, and thus degrades the time resolution of the excitation patterns. This trade-off is also shown in [35] by



**Figure 4.** Time resolution term  $dh/dt$  for low and high frequencies using a Kaiser window: The time resolution is non-uniform over both time and frequency.

<sup>2</sup> Note that although Eqs (8), (9) (and thereafter) explicitly show the frequency dependency in  $h(t,f)$ , in our implementation of  $h(t,f)$  and its derivative,  $f$  is treated as a constant, and only  $t$  is the variable.



neurophysiological experiments and in [36] by the gap-in-noise detection experiments, which provide evidence that human subjects are able to detect shorter gaps in a narrow band of noise when the noise bands are centered at higher frequencies. Despite of this property of human hearing (high time resolution for high frequencies), it's not clear whether this effect can be exploited for improving ASR. Our work provides one way to investigate this effect in features used for ASR.

Although the principles and forms for frequency and time warping have been presented, the magnitude of the power spectrum on the perceptual scale is the same as for the physical domain. To better represent perceptual magnitudes, the power spectrum should also be nonlinearly scaled. This nonlinear scaling is represented by the function  $a$ , typically logarithmic or power function with a low exponent such as  $1/15$ . Eq. (2) can be rewritten in terms of  $t$  and  $f$  by substituting in Eqs. (3), (4), (5), (8), (9):

$$Feat(i, j) = \int_{t=-1/2}^{1/2} \int_{f=0}^1 a(X(t, f)) \cdot \cos(\pi i g(f)) \frac{dg(f)}{df} \cdot \cos(\pi j h(t, f)) \frac{dh(t, f)}{dt} df dt \quad (9)$$

Eq. (10) can be written using modified basis vectors over frequency  $f$  as:

$$\varphi_i(f) = \cos(\pi i g(f)) \frac{dg(f)}{df}, \quad (10)$$

$$0 \leq i \leq N - 1.$$

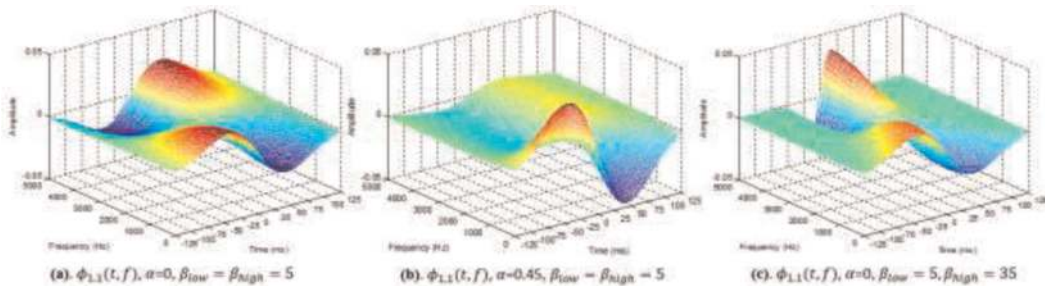
and modified frequency-dependent basis vectors over time  $t$  as:

$$\psi_j(t, f) = \cos(\pi j h(t, f)) \frac{dh(t, f)}{dt}, \quad (11)$$

$$0 \leq j \leq M - 1.$$

Using the basis vectors from Eqs. (11), (12), Eq. (10) can be expressed as:

$$Feat(i, j) = \int_{t=-1/2}^{1/2} \int_{f=0}^1 a(X(t, f)) \cdot \phi_{ij}(t, f) df dt. \quad (12)$$



**Figure 5.** Two-dimensional basis vector  $\phi_{1,1}(t, f)$  with bilinear frequency warping  $g(f)$  and a Kaiser window for  $dh(t, f)/dt$ .  $\alpha$  is the frequency-warping coefficient as in Eq. (7), and  $\beta_{low}$ ,  $\beta_{high}$  are the time-warping factors for low and high frequencies, respectively.

where the two-dimensional basis vectors  $\phi_{i,j}(t,f)$  are the product of the basis vectors given in Eqs. (11) and (12).

In **Figure 5**, the two-dimensional basis vector  $\phi_{1,1}(t,f)$  is plotted with bilinear frequency warping  $g(f)$  and a Kaiser window for the  $dh(t,f)/dt$  term. Panels (a) and (b) are based on the same time-warping factor  $\beta = 5$  for all frequencies, and only the frequency-warping factor  $\alpha$  is varied. Compared with the linear frequency scale ( $\alpha = 0$ ) in panel (a), the basis vector becomes more sharply peaked at low frequencies in panel (b) as higher frequency resolution is incorporated through a larger warping factor  $\alpha = 0.45$ . Panel (c) uses increasing time warping as frequency increases. The Kaiser window  $\beta$  value is linearly interpolated between  $\beta_{low}$  and  $\beta_{high}$ . The higher time resolution for high frequencies makes the basis vectors more concentrated near the center of the block.

Another option for the cosines used as the starting point for the two-dimensional basis vectors is to use a Gabor filterbank. As described in the work of [26, 27, 37], Gabor filtering is performed as a two-dimensional correlation between the Gabor filterbank and the perceptual time-frequency plane  $(t', f')$ . Each Gabor filter is defined using the product of a two-dimensional Gaussian envelope and a complex exponential function over a localized region in the time-frequency plane. Directionality is the most apparent difference between Gabor filter approach and the cosine expansion used in this chapter. Gabor filters can be adjusted toward any direction whereas the cosine transform only represents modulation of the spectrum along the vertical and horizontal axes. The deeper reason for this difference is that the Gabor approach and the method presented in this chapter are motivated by different considerations. The power spectrum directionality property of Gabor features stems from the response of neurons to combinations of spectral-temporal modulation frequencies in the spectral-temporal receptive field [38]. In contrast, the proposed framework is intended to model the trade-off between time and frequency resolution of the peripheral auditory system. However, it is possible to modify the proposed front end to incorporate the directionality of spectral-temporal patterns in a way similar to the Gabor filterbank. In prior work [39], this was achieved by rotating the 2-D cosine basis vectors by various angles.

### 3. Implementation

The 2-D integral in Eq. (10) can be implemented in a variety of ways, as discussed below. As mentioned previously, integrations are computed using sums and vector inner products between basis vectors and the sampled time-frequency plane.

#### 3.1. DCTC/DCSC method

The first version of the implementation is based on frequency-independent time warping; i.e. the time warping  $h(t,f)$  is simplified to  $h(t)$  for all frequencies. In this case, integrating in any order (first over  $f$  and then over  $t$  or the reverse) is equivalent. Conventionally, frequency integration is performed first, which generates a set of intermediate static features<sup>3</sup> called Discrete Cosine Transform Coefficients (DCTCs):

<sup>3</sup> Note that the term “static features” refers only to the outputs of the DCTC step. As mentioned in the beginning of Section II, the final outputs are the spectral-temporal features, which are computed by another integration over the time sequence of these “static” features.

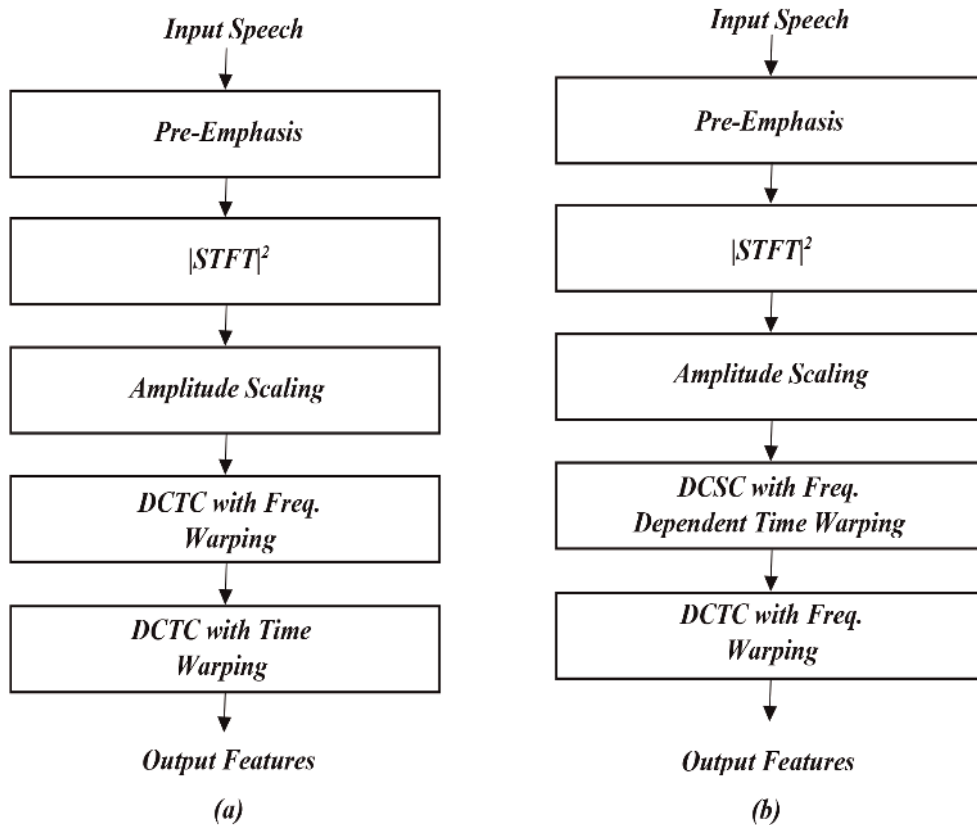
$$DCTC(i) = \int_{f=0}^1 a(X(t, f)) \cdot \varphi_i(f) df, \quad (13)$$

where  $\varphi_i(f)$  is the  $i$ th static basis vector as defined in Eq. (11). Then the trajectories of these DCTCs are encoded by integrations over time, yielding a set of features referred to as Discrete Cosine Series Coefficients (DCSCs):

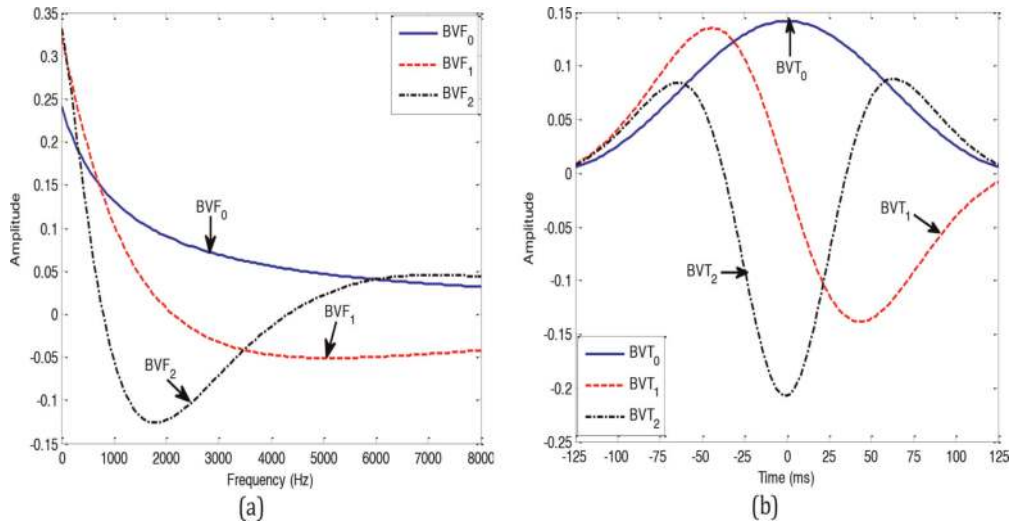
$$DCSC(i, j) = \int_{t=-1/2}^{1/2} DCTC(i) \cdot \psi_j(t) dt, \quad (14)$$

where  $\psi_j(t)$  is the  $j$ th basis vector over time, as defined in Eq. (12), but without dependence on  $f$ . These DCSC 2-D features, arranged as a 1-D feature vector, are the input to a recognizer. This implementation is depicted in **Figure 6(a)**. **Figure 7** is a plot of the first three DCTC and DCSC basis vectors, using a Mel-shape frequency warping and a Kaiser window with  $\beta = 5$  for (derivative of) time warping. The zeroth order terms represent the form of the spectral/temporal resolution.

Unlike some other front ends, such as RASTA [40], TRAPS [41], as well as the Gabor method mentioned previously, for which modulation frequencies are a key concept, the proposed DCTC and DCSC method does not explicitly use this concept.



**Figure 6.** Two implementations of the proposed front end: (a) the DCTC/DCSC implementation in which DCTCs are computed first followed by DCSCs. The time warping in the DCSC basis vectors is uniform for all frequencies. (b) the DCSC/DCTC implementation in which a set of DCSCs are obtained first followed by DCTCs. This implementation enables frequency-dependency in the DCSC basis vectors.



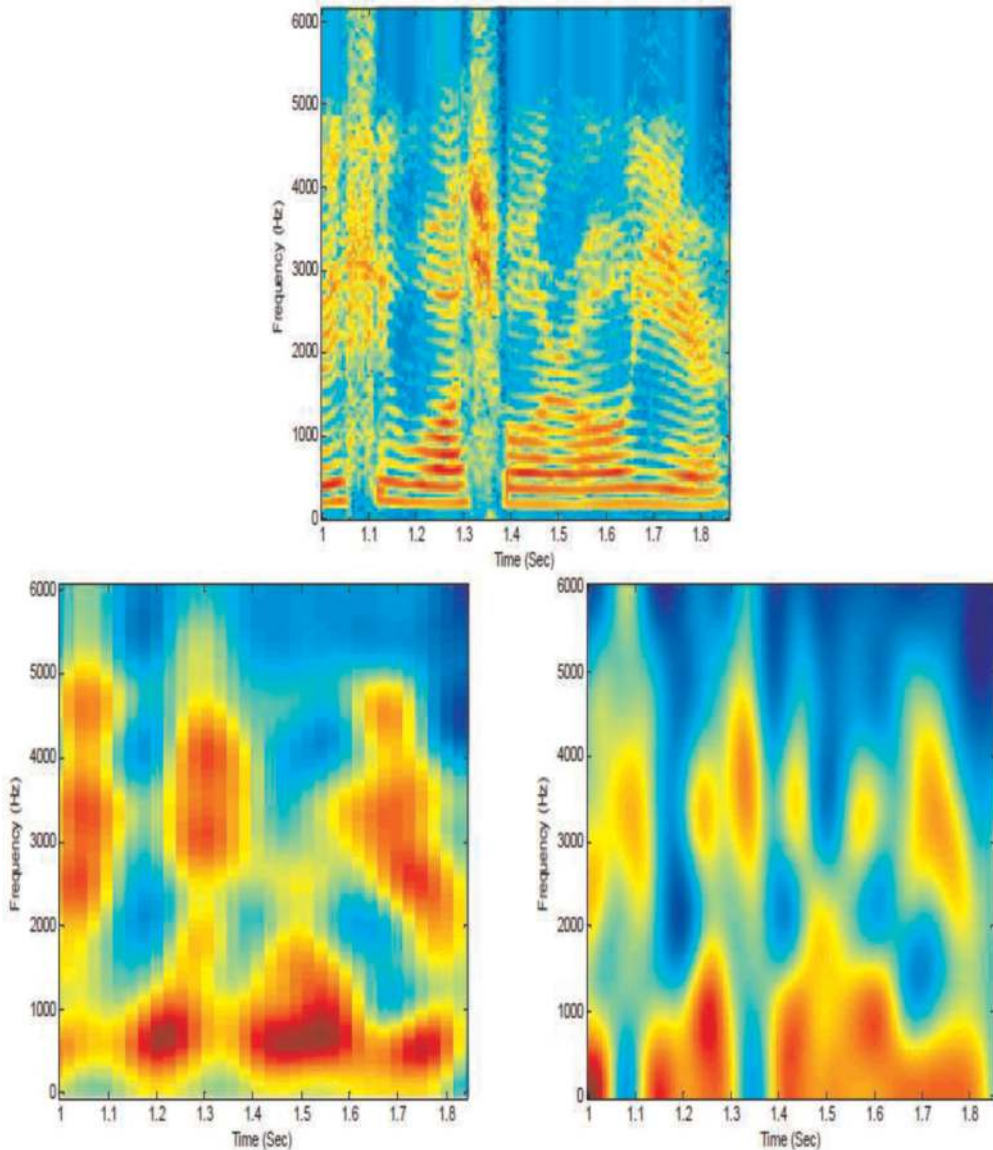
**Figure 7.**

The first three DCTC (a) and DCSC (b) basis vectors: A Mel-shape and a cumulative Kaiser window are used for frequency and time warping respectively.

The DCSC basis vectors act as non-causal FIR low pass temporal filters of spectral dynamics. Similarly, the DCTCs can also be viewed as low pass filtering of the power spectrum. Parameters used in the DCTC/DCSC implementation can be varied to examine the trade-offs between spectral and temporal resolution. The trade-off between spectral and temporal resolution considered here is different than the auditory time-frequency resolution built into the warping of the basis vectors as presented previously. Here, based on the filtering point of view, the parameters determine how much detail of the static spectrum and dynamic trajectory is preserved after the low pass filtering. The time-frequency resolution represented by the derivatives of the warping (which also cause a trade-off effect) is an intrinsic property of human hearing. As mentioned, the proposed DCTC/DCSC front end can be tuned to emphasize either side of the overall spectral or temporal resolution. For increased emphasis on the spectral information, a long frame length and a relatively large number of DCTCs should be used, with a relatively small number of DCSCs computed from a long block length. For increased emphasis on time resolution, a short frame length and frame spacing should be used with a large number of DCSCs computed from a short block length.

**Figure 8** graphically illustrates this spectral-temporal trade-off. The top panel depicts the unprocessed spectrogram of a speech segment. Two spectrograms reconstructed from DCTC/DCSC terms are shown in the bottom panels<sup>4</sup>. The left one has high spectral resolution and low temporal resolution. It is rebuilt using 16 DCTCs, computed using 25 ms frames, a 10 ms frame spacing and 4 DCSCs with a block length of 50 frames (500 ms). The one in the right bottom panel has low spectral resolution but high temporal resolution. It is computed from 8 DCTCs, 5 ms frames spaced by 2 ms, and 6 DCSCs with a block length of 100 frames (200 ms). The low frequency components in both rebuilt

<sup>4</sup> Briefly, to rebuild the spectrum, the DCTCs and DCSCs are computed using orthonormal basis vectors, which can be obtained using Gram-Schmidt orthonormalization. Then the DCTCs of the center frame of a block are rebuilt first by multiplying the DCSCs by the transpose of the DCSC basis vector matrix and preserving only the center frame. Then the spectrum of this frame is rebuilt in a similar way by a matrix product using the transpose of the DCTC basis vector matrix.



**Figure 8.** Spectrogram of a speech segment (upper panel) and two rebuilt spectrograms: The bottom left one has high spectral resolution and low temporal resolution while the bottom right one has low spectral resolution but high temporal resolution.

spectrograms are represented with higher resolution than are the higher frequency components due to the Mel frequency warping. Comparing the two reconstructed spectrograms, the spectrogram in the left panel preserves more spectral details than does the spectrogram in the right panel. In contrast, the spectral dynamics are shown with more resolution in the right hand panel than are the spectral dynamics in the left pane.

### 3.2. DCSC/DCTC method

In the case of frequency-dependent time warping, the 2-D integration in Eq. (10) can be implemented by integrating over the time axis first followed by another integration over frequency. **Figure 6(b)** depicts the diagram of this configuration. In this case, Eq. (10) can be rearranged as:



$$Feat(i,j) = \left[ \int_{f=0}^1 \cos(\pi ig(f)) \frac{dg(f)}{df} \left[ \int_{t=-1/2}^{1/2} a(X(t,f)) \cdot \cos\left(\pi jh(t,f)\right) \frac{dh(t,f)}{dt} dt \right] df \right] \quad (15)$$

The inner integral defines a set of frequency-dependent DCSCs,

$$DCSC(j,f) = \int_{t=-1/2}^{1/2} a(X(t,f)) \cdot \psi_j(t,f) dt, \quad (16)$$

where  $\psi_j(t,f)$  is the  $j$ th DCSC basis vector for frequency  $f$ , as defined in Eq.(12). Then, the integral over frequency computes the DCTCs, which yields the final features

$$Feat(i,j) = DCTC(i,j) = \int_{f=0}^1 DCSC(j,f) \cdot \varphi_i(f) df, \quad (17)$$

where  $\varphi_i(f)$  is the  $i$ th DCTC basis vector as in Eq. (11).

### 3.3. Unified framework

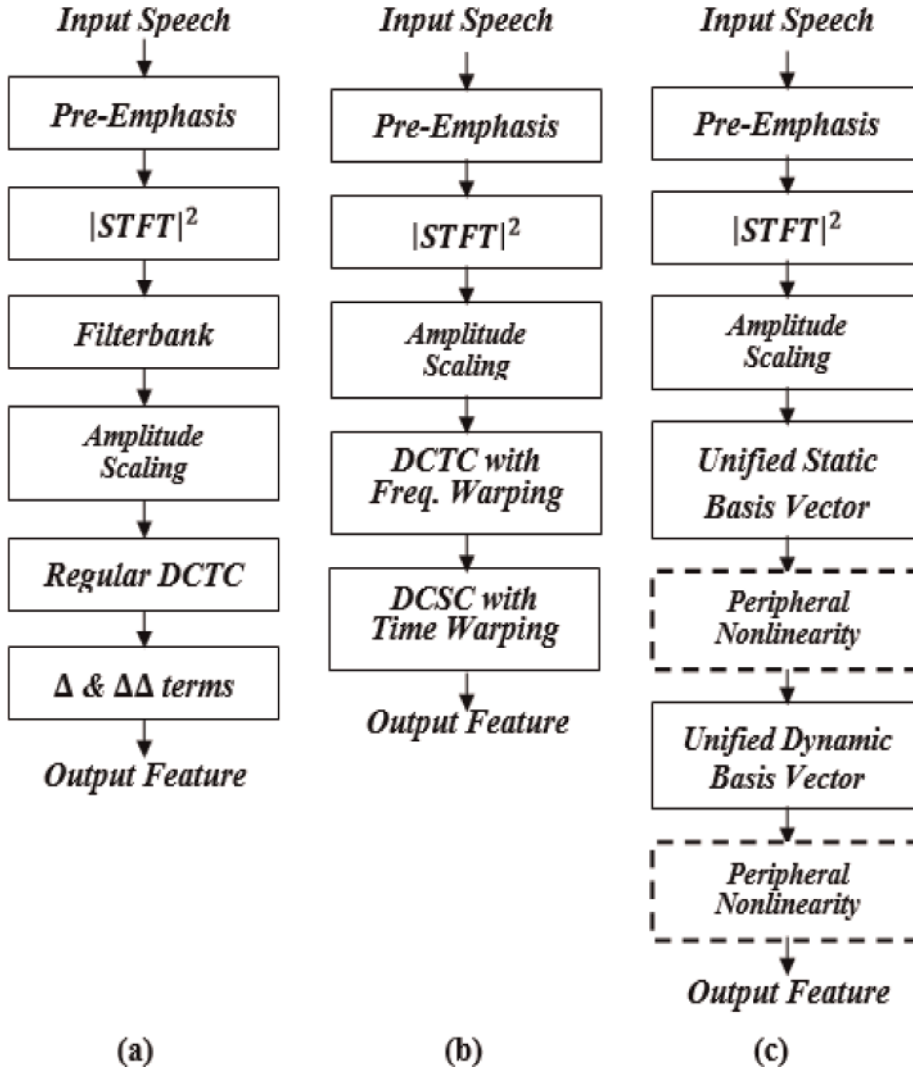
As mentioned in Section 1, the DCTC/DCSC structure proposed in this chapter can be viewed as a unified framework which incorporates the filterbank implementation of the frequency warping as well as the conventional delta and acceleration dynamic features. To illustrate this unified viewpoint, a comparison of the “standard” MFCC front end and the DCTC/DCSC front end is presented in **Figure 9**.

In the filterbank-based front end, the frequency warping is performed by a group of auditory filters, and followed by a “regular” DCT transform with the term “regular” referring to sampled versions of half cosine basis vectors (in contrast to the basis vectors proposed in the previous sections). Specifically, the regular DCT transform is given by:

$$c(i) = \sqrt{\frac{2}{Q}} \sum_{j=1}^Q a(P(j)) \cos\left(\frac{\pi i}{Q} (j - 0.5)\right), \quad (18)$$

where  $c(i)$  is the  $i$ th DCT coefficient,  $Q$  is the total number of filter channels,  $P(j)$  is the output power of the  $j$ th channel, and  $a(\cdot)$  is the amplitude scaling function. The terms  $\cos\left(\frac{\pi i}{Q} (j - 0.5)\right)$  are the unmodified cosine basis vectors.

In prior work [42], it was experimentally verified that the nonlinear amplitude scaling in the filterbank based front end can be moved to immediately before the filterbank without degrading ASR performance (i.e. swap the position of the filterbank block and the amplitude scaling block in **Figure 9(a)**). Then the filterbank weights can be combined with the unmodified cosine basis vectors by a simple matrix multiplication. Mathematically, suppose each row of the matrix  $\mathbf{W}$  contains the magnitude response of a filterbank channel (i.e. if 26 channels are used with 128 FFT samples for each channel,  $\mathbf{W}$  is a 26 by 128 matrix), and each row of the matrix  $\mathbf{BVF}_{reg}$  contains the 12 unmodified cosine basis vectors,  $\mathbf{BVF}_{reg}$  is a 12 by 26 matrix). A set of unified static basis vectors  $\mathbf{BVF}_{uni}$ , which incorporate the filterbank, can be formed by a matrix multiplication:



**Figure 9.** Block diagrams of the filter bank front end (a), the DCTC/DCSC front end (b) and a unified framework (c) of (a) and (b) dashed blocks (— —) are optional.

$$BVF_{uni} = BVF_{reg} \cdot W \quad (19)$$

In the proposed DCTC/DCSC case,  $BVF_{uni}$  is simply the matrix of the basis vectors  $\varphi_i(f)$  defined in Eq. (11) with each row containing one such basis vector. Thus, with the unified static basis vectors, the static features in the filterbank front end and the DCTC/DCSC front end can be obtained using the same mathematical framework. The only difference lies in how their basis vectors are computed. Specifically, if the matrix  $\mathbf{X}$  represents the power spectrum of a block of frames<sup>5</sup> for which each column is the magnitude squared STFT for a frame, the static features of this block for both the

<sup>5</sup> For consistency with the block processing in the computation of the dynamic features, the static feature computation also uses block notation here. When implemented, the static features are computed for the entire utterance once, and only the final features are computed block by block. That is, in the static feature step,  $\mathbf{X}$  represents the spectrum of the entire utterance, and in dynamic feature step in Eq. (22),  $\mathbf{X}$  denotes a block of frames.

filterbank front end and the DCTC/DCSC front end can be computed in a unified way as  $\mathbf{BVF}_{uni} \cdot a(\mathbf{X})$  where  $a(\mathbf{X})$  represents the amplitude scaling.

Similarly, the delta and higher order dynamics in the standard MFCC front end can also be computed by a summation of the static features over time, weighted by a set of dynamic basis vectors. From Eq. (1), to compute any  $n$ th order differential term, its basis vector with respect to the previous lower order terms (neglecting the constant denominator) is given by  $\mathbf{bv}_n = [-\theta_n, -\theta_n + 1, \dots, 0, 1, \dots, \theta_n]$  where  $\theta_n$  is the window length in Eq. (1). Considering  $\mathbf{bv}_n$  as a discrete signal with each element representing both the index and the amplitude (i.e.  $[-3, -2, -1, 0, 1, 2, 3]$  gives a signal whose magnitude is  $-3$  at index  $-3$ , and  $-2$  at index  $-2$ , etc.), then the  $n$ th order delta basis vector  $\mathbf{bvT}_n$  can be computed as

$$\mathbf{bvT}_n = \mathbf{bv}_1 \otimes \mathbf{bv}_2 \dots \otimes \mathbf{bv}_n. \quad (20)$$

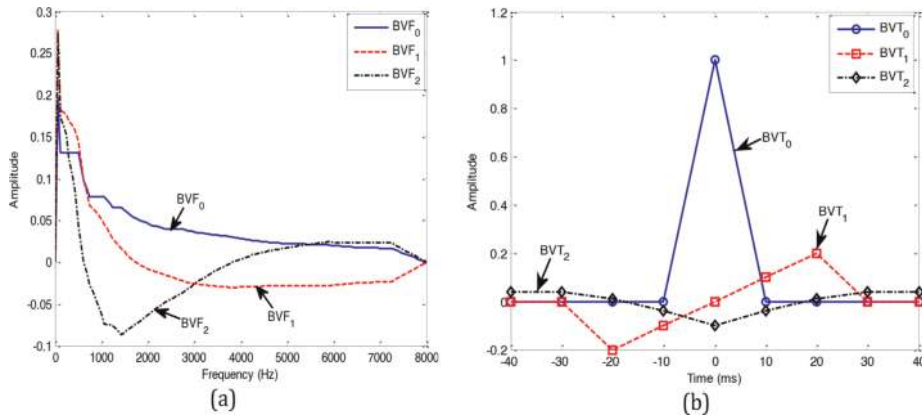
where  $\otimes$  is the convolution operator. Thus, a set of unified dynamic basis vectors  $\mathbf{BVT}_{uni}$  can be defined. In the case of the delta features, each row of  $\mathbf{BVT}_{uni}$  stores one dynamic basis vector of the form in Eq. (21) whereas in the proposed DCTC/DCSC front end, each row of  $\mathbf{BVT}_{uni}$  stores one DCSC basis vector as defined in Eq. (12). Hence, again the final output features  $\mathbf{F}$  for both the MFCC and the DCTC/DCSC methods can be written in a unified way:

$$\mathbf{F} = \mathbf{BVT}_{uni} \cdot [\mathbf{BVF}_{uni} \cdot a(\mathbf{X})]^T \quad (21)$$

**Figure 9(c)** is a block diagram of this unified framework. This diagram depicts the essence of the proposed speech features as well as similar features such as MFCCs. They are essentially a series of linear transformations of the spectrum scaled by an auditory nonlinearity with optional peripheral nonlinearities in between (dashed blocks in the diagram), such as the sigmoid-shaped functions given in [43, 44]. These nonlinearities generally improve the noise robustness of front ends. In this work, the linear transformations are represented by unified basis vectors. Filterbank-based features (such as MFCC or PLP) exert their impact on features by shaping the basis vectors implicitly. The unified basis vectors presented here determine the properties of a front end. Thus we have a common yardstick with which to analyze and compare front ends based on the properties of the unified basis vectors.

A basic comparison can be made between filterbank-based frontends such as the widely-used MFCCs and the proposed DCTC/DCSC front end by comparing their unified basis vectors. Although the MFCC front end and the DCTC/DCSC front end are derived differently, the unified framework shows the two approaches are the same, except that the basis vectors are different.

**Figure 10** is a plot of the first three unified static basis vectors underlying MFCC features (based on 26 Mel filters) and three unified temporal basis vectors used to compute the zeroth order, delta and acceleration terms. The unified basis vectors over frequency are not as “smooth” as the ones proposed here which are based on the continuous Mel-shape warping  $g(f)$ , as shown in **Figure 7(a)**. The “jagged” basis vectors ‘plotted in **Figure 10(a)** result from the quantization effect caused by the coarse sampling of the frequency axis by the filter bank. The unified temporal basis vectors, implicit in most current methods, estimate derivatives very approximately using a small number of samples. A comparison of the temporal basis vectors (see **Figures 7(b)** and **10(b)**) graphically illustrate that the standard delta/acceleration method uses only a few central terms in each block whereas in the proposed method,



**Figure 10.** The first three unified static basis vectors resulting from 26 Mel filters (a) and the first three unified dynamic basis vectors of the delta method (b).

the incorporation of non-uniform time resolution result in long “smooth” basis vectors emphasizing the center of the block but extending to the ends of the block. A comparison of both panels of **Figure 7** with both panels of **Figure 10** clearly illustrate the more continuous nature of the temporal basis vectors for the proposed method versus the implicit basis vectors corresponding to delta and acceleration terms. This suggests that the proposed DCSC basis vectors may represent spectral dynamics with more accuracy and resolution than is the case for delta/acceleration method.

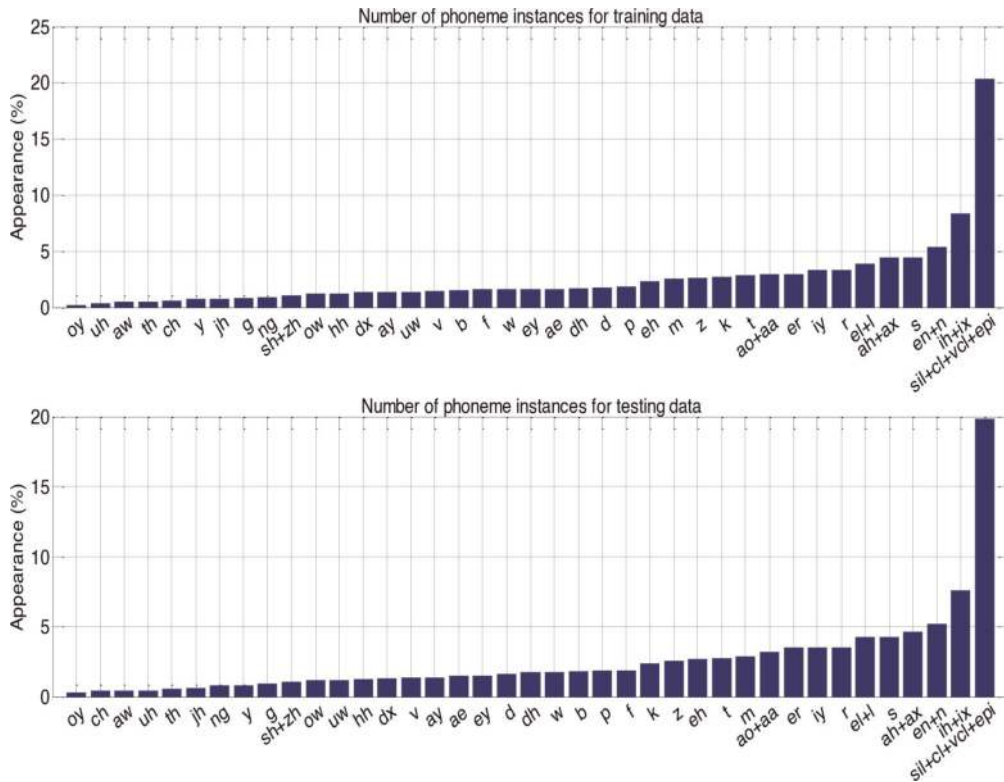
## 4. Experimental evaluation

### 4.1 Experimental configuration

A comprehensive suite of ASR tests for various conditions and parameter settings was performed to evaluate the effectiveness of the spectral-temporal DCTC/DCSC features and to investigate trade-offs in time and frequency resolution as that affects ASR performance. All experiments reported in this chapter are for phone recognition, with monophone models using the HTK 3.4 HMM/GMM recognizer [45]. Except for one set of evaluation experiments described below, all experiments use the TIMIT database [46]. As is typically done with this database, 3696 utterances (462 speakers, eight sentences/speaker, approximately 236 minutes) with SA sentences removed were used for training. The TIMIT database document [46] suggests using 1344 utterances (168 speakers, eight sentences/speaker, approximately 86 minutes) for testing. However, since various parameters in the proposed front end needed to be tuned both for performance optimization and for exploring the effects on the time-frequency properties, a development set (DEV set) was needed. Thus, 672 utterances from the original test set were randomly chosen for this purpose, and the remaining 672 utterances were used as the evaluation set (EVAL set). Also, as recommended in [47], the original set of 61 labeled phones was collapsed to 48 phones to create 48 phone models with a further reduction to 39 phone categories for scoring. Some similar phones were merged to create the 39 categories: For convenient reference, the reduction from 61 to 48 phones and further from 48 to 39 phones (shaded) is presented in **Table 1**, and a frequency count of the 39 phones for the training and the original test sets is shown in **Figure 11**. All HMM acoustic models had three emitting

TIMIT Phone	Reduced Phone	TIMIT Phone	Reduced Phone	TIMIT Phone	Reduced Phone
oy	oy	v	v	er axr	er
uh	uh	b	b	iy	iy
aw	aw	f	f	r	r
th	th	w	w	el	el
ch	ch	ey	ey	l	l
y	y	ae	ae	ah	ah
jh	jh	dh	dh	ax-h ax	ax
g	g	d	d	s	s
ng eng	ng	p	p	en	en
sh	sh	eh	eh	n nx	n
zh	zh	m em	m	ih	ih
ow	ow	z	z	ix	ix
hh vv	hh	k	k	#h pau	sil
dx	dx	t	t	pcl tcl kcl qcl	cl
ay	ay	ao	ao	bcl dcl gcl	vcl
uw ux	uw	aa	aa	epi	epi

**Table 1.** 61 TIMIT phones, reduced to 48 for training, and 39 categories (shaded) for testing.



**Figure 11.** A frequency count of the 39 TIMIT phone categories.



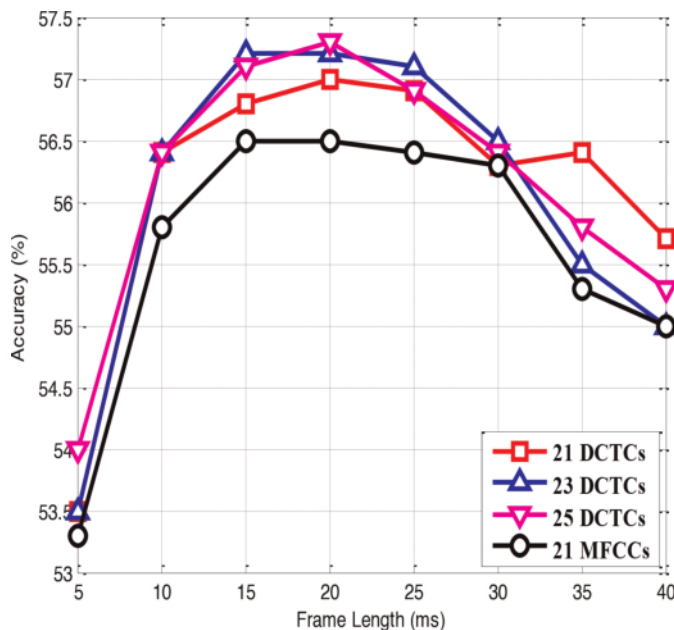
hidden states. A bigram language model was used based on phone bigram frequencies in the training set.

Since invariability is also crucial for “good” features, which means that the optimal front end parameters experimentally tuned from a DEV set should work approximately equally well on different independent evaluation sets without the need of re-tuning the parameters, an independent phone recognition task was also conducted using the Chinese Mandarin 863 Annotated 4 Regional Accent Speech Corpus (RASC863) [48]. The phonetically transcribed portion of this database was used for this work, which includes 20 speakers, each uttering 110 phonetically balanced sentences. Due to the much smaller number of speakers than for TIMIT, approximately 70% of the total set of 2200 utterances from all of the 20 speakers were used for training (1540 sentences, approximately 77 sentences/speaker and 224 minutes), and the remaining 30% were used for evaluation (660 sentences, 33 sentences/speaker, 96 minutes). Fifty-nine Chinese base phones (without considering tone information) were trained and evaluated on the evaluation set against the baseline directly using the optimal parameters obtained from the TIMIT experiments.

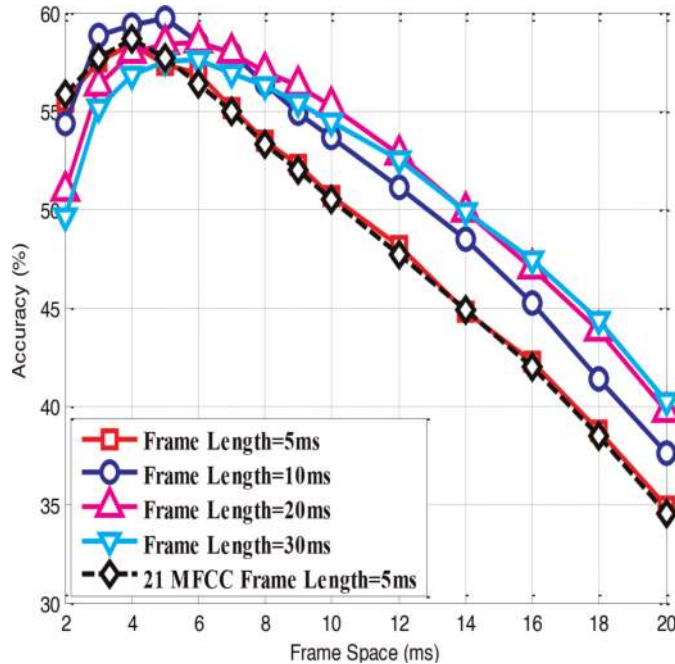
Processing begins with a complex pole pair IIR pre-emphasis filter:

$$y[n] = x[n] - 0.95x[n - 1] + 0.494y[n - 1] - 0.64y[n - 2] \quad (22)$$

This second order filter has a peak near 3200 Hz and is a reasonably good match to the inverse of the equal-loudness contour for human hearing. In our previous work [49], it was found that this filter results in slightly higher ASR accuracy than is obtained with the more typically-used first-order one zero pre-emphasis. All speech passages were then divided into overlapping windowed frames (Kaiser window with  $\beta$  of 6, similar to a Hamming window). A 512 point FFT of each frame was computed, and log magnitudes computed, for a frequency range of 100 Hz to 7000 Hz. For each frame, log magnitudes were “floor” clipped at 40 dB below the largest spectral magnitude in each frame. In previous work [50], this simple floor was found to improve



**Figure 12.** Phone recognition accuracy as function of frame length using 21, 23, and 25 DCTCs.



**Figure 13.**  
Effect of frame length and frame space on phone recognition accuracy for 21 DCTCs.

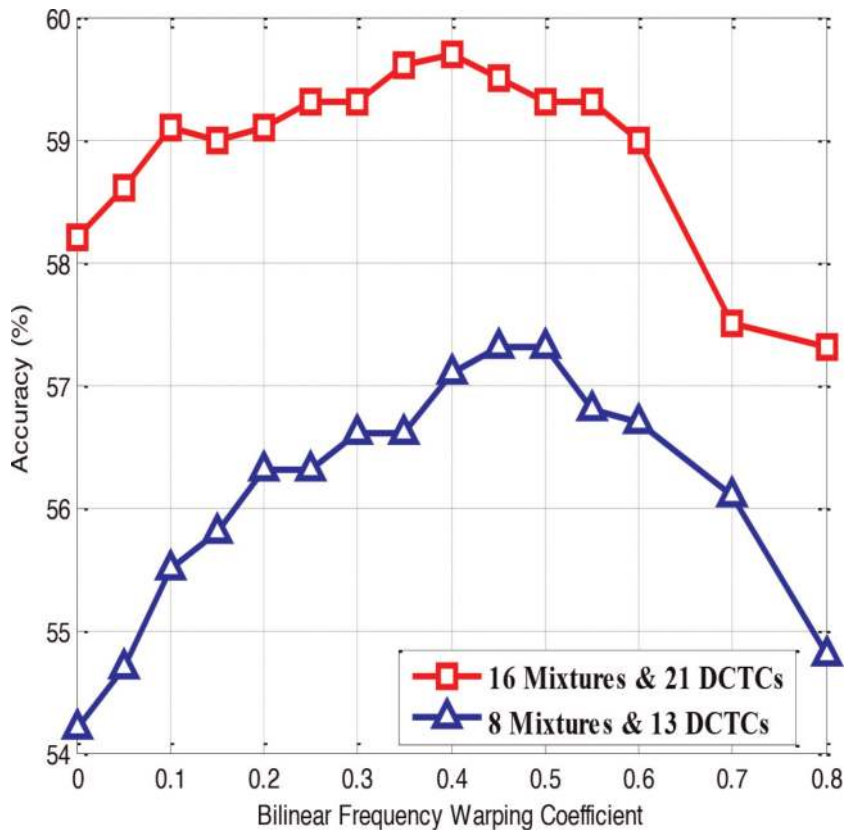
ASR accuracy by a small amount, especially for noisy speech. In summary, each sentence was converted to a matrix of spectral values which were then further processed by the DCTC/DCSC methods proposed in this chapter.

## 4.2 TIMIT DEV set parameter optimization

### 4.2.1 Experiment set 1: DCTC features only (static features)

For the experiments reported in this section, DCTC features only were computed. The goal was to experimentally evaluate how frame length, frame space, number of DCTCs, and type and degree of frequency warping affects ASR accuracy. Not all combinations of parameter values are presented in the results due to the very large number of combinations. Rather, most of the parameter values were fixed at what appeared to be the best values based on pilot experiments, and then a subset of parameter values was varied and performance evaluated.

*Experiment A1— Spectral resolution issues for DCTCs:* The goal was to examine spectral resolution effects on ASR performance as determined by frame length and number of DCTCs. The frame space was fixed at 8 ms. Mel frequency warping (bilinear warping with a coefficient of .45) and 16 mixture GMM/HMMs were used. The spectrum of each frame was represented with 9 to 26 DCTCs. Frame length ranged from 5 ms to 40 ms. ASR accuracy ranges from approximately 49–57% in these tests. **Figure 12** depicts ASR accuracy using 21, 23, and 25 DCTCs as a function of frame length. It also contains the static MFCC baseline results using 26 filters, 21 DCTCs, again with the frame space fixed at 8 ms. The absolute best accuracy (57.3%) was obtained with 20 ms frames and 25 DCTCs. However, the increase in performance for more than 19 DCTCs is minimal, typically less than 0.5%. Frame lengths of 15 ms to 30 ms result in fairly similar ASR accuracies.



**Figure 14.**

Phone recognition accuracy as function of frequency warping for two cases: The standard Mel warping, i.e.  $g(f) = 2595 \log(1 + f/700)$ , was used as a baseline, and results were within 0.1% of bilinear warping with a coefficient of 0.45 in both cases.

*Experiment A2—Time resolution effect for DCTCs:* To investigate the role of time resolution on frame-based speech features, the feature “sampling rate” was varied by changing the frame spacing from 2 ms to 20 ms. Since the time resolution is also affected by frame length, 4 frame lengths (5, 10, 20, and 30 ms) were evaluated. 21 DCTCs were used for all tests. Other parameters were the same as for Experiment A1. The baseline in this experiment is the case of static MFCCs with frame length fixed at 5 ms. Results are shown in **Figure 13**.

Results vary from 34.9% (5 ms frames, 20 ms apart) to 59.7% (10 ms frames, 5 ms apart). Phonetic recognition accuracy degrades when the frame space is too large, especially for shorter frame lengths. The best performance for each frame length varies from 57.6% to 59.7%. As might be expected, the highest accuracy is obtained with short frame spaces and short frame lengths—that is high time resolution. However, unexpectedly, accuracy degrades when the frame space is too short. We hypothesize that oversampling of features is problematic for the HMM recognizer, due to the high correlation of features when frames are very closely-spaced.

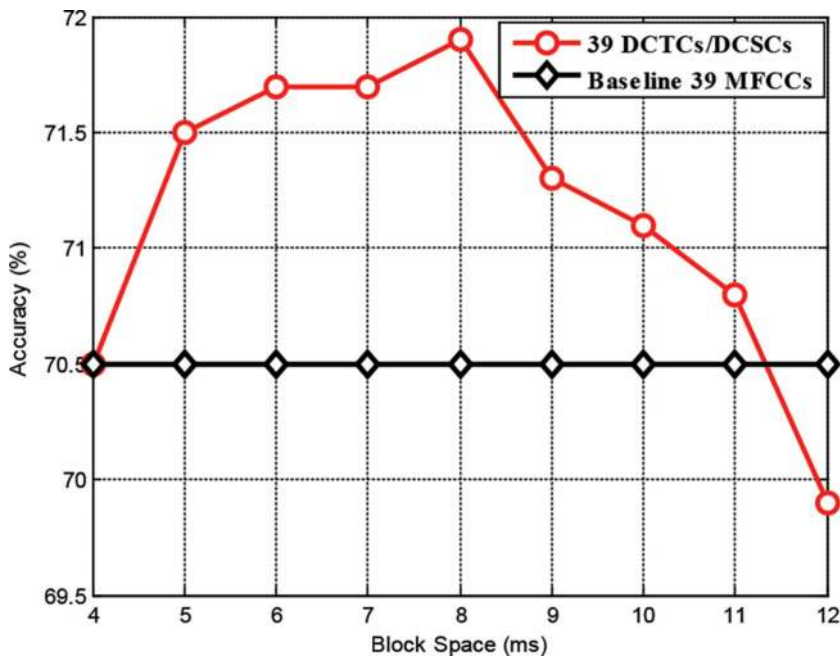
*Experiment A3—Effect of frequency warping on DCTC features:* To evaluate frequency warping, bilinear frequency warping was used as in Eq. (7) with a single parameter  $\alpha$  controlling the shape of the nonlinearity for the frequency warping. Bilinear warping with a coefficient of 0.45 closely approximates Mel warping, whereas a coefficient in the range of 0.5 to 0.57 approximates Bark warping [32].

Since pilot experiments showed that the effects of frequency warping depend on the number of DCTC features and the number of HMM mixtures in the recognizer, these experiments were performed for two cases—13 DCTCs with 8 mixture HMMs; 21 DCTCs with 16 mixture HMMs. 10 ms frames spaced 5 ms apart were used in all cases. Results are plotted in **Figure 14** as the warping coefficient varies from 0 (no warping) to 0.8 (over warped).

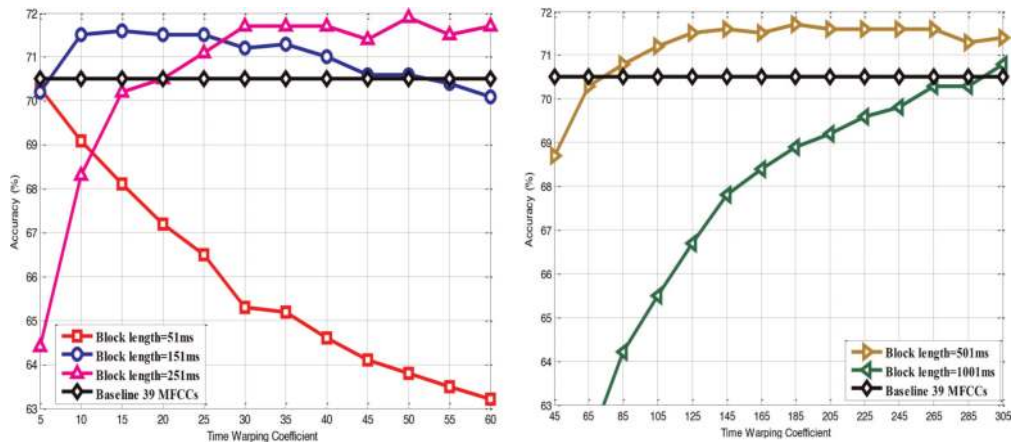
The effect of warping is more apparent for the 8 mixture case than for the 16 mixture case. The overall best warping values found were .4 and .45 (most similar to Mel warping). For the 8 mixture/13 DCTC cases, the best warping of .45 resulted in a 3% accuracy improvement over the no warping case. For the 16 mixture/21 DCTC case, the best warping of .4 yielded a 1.5% accuracy improvement over the no warping case. The “standard” Mel warping, as proposed by O’Shaughnessy [31], was also evaluated as a baseline, and the result was within 0.1% of the result obtained using a bilinear warping coefficient of 0.45 for both 13 DCTCs/8 mixtures and 21 DCTCs/16 mixtures.

4.2.2 Experiment set 2: Dynamic features (DCTCs and DCSCs)

In these experiments, a myriad of parameters believed to be significant for DCTC/DCSC features which represent spectral-temporal characteristics in a block of frames centered on each frame were varied. These parameters include number of DCTCs/DCSCs, frame length/space, frequency/time-warping coefficients, and block length/space. Not all combinations of parameters were tested due to both the very large number of cases and the assumption that many of the variations would have much effect on ASR accuracy. Based on pilot experiments and the results reported previously for experiments B1, B2, and B3, many of these parameters were either fixed to a



**Figure 15.** Phone recognition accuracy of 39 DCTCs/DCSCs as function of block space with block length fixed at 251 ms: The 39 MFCC features produce a baseline of 70.5% (block space fixed at 8 ms).



**Figure 16.** Phone recognition accuracy as function of time-warping factor for different block lengths with a fixed block spacing of 8 ms: The baseline 39 MFCC case is also depicted.

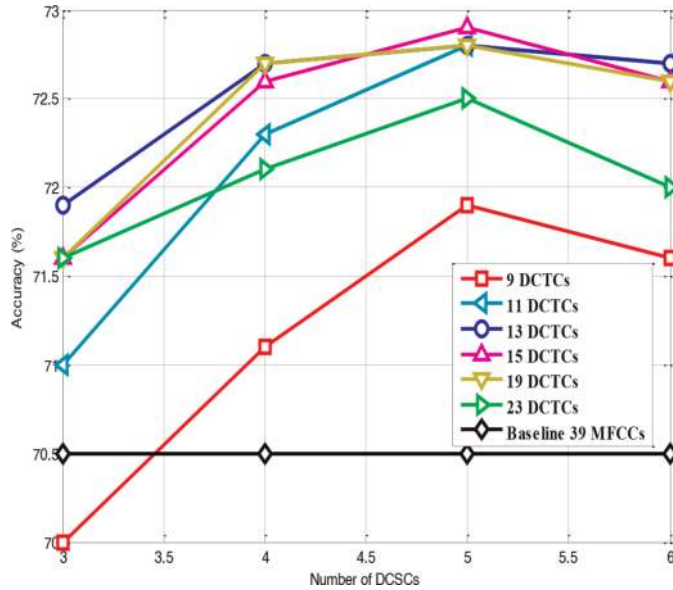
single value or varied over a small range. The other parameters were varied and performance evaluated. 32 HMM mixtures were used due to the large dimensionality of the feature space.

*Experiment B1—39 feature (13 DCTCs/3 DCSCs) experiments:* Since 39 MFCC features are often used for ASR systems, the first experiments were performed with 39 features—13 DCTCs/3 DCSCs. The 39 MFCC feature case was the baseline. In this these experiments, the block length was fixed at 251 ms, and the effect of block spacing, which was also the feature “sampling” rate, was varied from 4 ms to 12 ms, with results depicted in **Figure 15**. The frame length/space was fixed at 10 ms/1 ms, and bilinear frequency warping (coefficient of 0.45) was used. The time-warping coefficient was 50 using a Kaiser window. The effect of block space on ASR accuracy varies approximately 2% from the lowest (12 ms) case to the highest (8 ms) case.

*Experiment B2—39 features (13 DCTCs, 3 DCSCs), block length and time-warping effects (auditory time resolution):* The objective of this experiment set was to examine the role of block length and time warping in representing the feature trajectories. These two parameters are closely related to the auditory time resolution of feature trajectories: A longer block length gives the ability to represent lower temporal modulation frequencies. A higher time-warping factor corresponds to higher time resolution in the central portion of a segment. To study these effects, five block lengths were used (51, 151, 251, 501, 1001 ms) with block spacing fixed at 8 ms. The time-warping factor of a Kaiser window was varied from 5 to 60 for the 51, 151, and 251 ms cases in steps of 5, and it was varied from 45 to 305 for the 501 and 1001 ms cases with steps of 20. The parameters for static features were identical to those in Experiment B1. Results are depicted in **Figure 16** again with a baseline of 39 MFCCs.

The highest accuracy of 71.9% was obtained with a time-warping coefficient of 50 using a block length of 251 ms. Results suggest that the block length and time warping are closely related to each other. As the block length increases, a larger time warping is required to achieve better performance, and a moderately long block length, such as 251 ms, which incorporates informative contextual information for each sample instant, provides the best result. However, very long contexts, such as 501 and 1001 ms, do not improve the performance and require very large-time-warping values. This shows that the spectral contexts too far from the current “observation point” do not provide much useful information, but can be suppressed by a large





**Figure 17.** Phone recognition accuracy as function of combinations of DCTCs and DCSCs.

time-warping factor, which emphasizes the useful information within a much shorter range surrounding the block center. Also, a long block length greatly increases the computations for each block (the number of multiplications in the vector inner product for computing the integration). Based on these considerations, 251 ms is considered the best value for the block length.

*Experiment B3—Overall spectral-temporal effect:* Phonetic recognition accuracy was evaluated with a variety of DCTC numbers (9 to 23 in steps of 2) and for a variety of DCSC numbers (3, 4, 5, and 6). These combinations were used to examine the trade-off between spectral and temporal resolution. Other parameters were selected to match the best parameter settings from earlier experiments (frame length/space of 10 ms/1 ms, bilinear frequency warping with coefficient of 0.4 for cases with 15 or more DCTCs and 0.45 for cases with fewer than 15 DCTCs, 251 ms/8 ms block length/space, and time warping of 50). Results are shown in **Figure 17**.

First, as the number of DCTCs increases beyond about 15, the performance begins to decrease. The number of DCSCs has a similar effect. Also, when a relatively small number of DCTCs were used, i.e. less overall spectral resolution, the performance increases relatively quickly as more DCSCs are used, i.e. more overall time resolution, as can be seen in the 9 and 11 DCTC cases (2% improvement from 3 DCSCs to 5 DCSCs using 9 DCTCs). However, the performance improves more slowly with more DCSCs when a relatively large number of DCTCs is deployed (less than 1% increment using 23 DCTCs). This observation shows the trade-off between the overall spectral and temporal resolution. The optimal “balance point” was obtained using 15 DCTCs and 5 DCSCs which produced 72.9% accuracy.

### 4.3 Independent EVAL set results and invariability

Based on the results from the TIMIT DEV set, a subset of parameters was further optimized. Two optimal parameter sets for a small feature set (27 features) and a large feature set (75 features) were obtained respectively. Also, the number of

GMM mixtures for each feature set was also optimized using the TIMIT DEV set. After these “final” optimizations were performed, to verify the generality of the tuned front end parameters, two EVAL phone recognition tasks were conducted with different data, as mentioned previously. The EVAL sets were the TIMIT EVAL set and the RASC863 Chinese Mandarin EVAL set. For the Chinese phone recognition task, the number of GMM mixtures was reduced due to the lower amount of available training data and the greater number of phone models to be trained. The best parameter values and the EVAL results are reported in **Tables 2 to 5**. In these tables, “BIG\_REC” refers to the results using the optimal number of GMMs, indicating the best accuracy achieved by a high order HMM recognizer. In addition, the accuracy for the training set in each case is also reported, which shows an ideal upper bound of the recognizer performance if the training data completely represents the test data.

It can be seen from these results that the proposed DCTC/DCSC method achieves generally better performance than the baseline MFCC for independent EVAL sets. In addition, to further examine the feature invariability of the DCTC/DCSC front end, for the Chinese phone recognition task, the parameter values based on the TIMIT DEV set were varied and re-evaluated. These tests showed (results not given here) that the parameter values for best performance did not change, which meant that the parameter values determined from the TIMIT DEV applicable to an entirely different database in a vastly different language.

*Experiment C1—DCTC/DCSC small feature set evaluation performance:* The optimum settings for a small feature set are summarized in **Table 2**. Accuracies on the EVAL sets are reported in **Table 3**.

*Experiment C2—DCTC/DCSC large feature set evaluation performance:* The optimum settings for a large feature set are summarized in **Table 4**, and accuracies on the EVAL sets are reported in **Table 5**.

#### 4.4 Unified framework explanation and statistical significance tests

As mentioned in Section 3 and in previous work [42], since the step of the amplitude scaling can be moved to immediately before the filterbank, the filterbank weights can be merged with the unwarped regular DCT basis vectors by a simple matrix product. Similarly, the delta and higher order acceleration dynamic terms can also be computed in a basis vector form. Thus, the proposed DCTC/DCSC front end and more

Parameter	Value
Frame Length	8 ms
Frame Spacing	1 ms
Frequency Warping $g(f)$	Bilinear, $\alpha = 0.45$
Number of DCTCs	9
Number of DCSCs	3
Frames per Block	251 ms (251 frames)
Block Spacing	7 ms (7 frames)
Time Warping ( $dh/dt$ term)	Kaiser window, $\beta = 50$

**Table 2.**  
*Optimum parameter settings for small feature set.*

TIMIT Database	Number of HMM mixtures	EVAL Accuracy (%)	Training Accuracy (%)
Baseline 27 MFCCs	16	66.7	70.0
DCTC/DCSC	16	68.9	72.1
BIG_REC Baseline 27 MFCCs	80	69.2	79.2
BIG_REC DCTC/DCSC	80	71.3	81.2
RASC 863 Database	Number of HMM mixtures	EVAL Accuracy (%)	Training Accuracy (%)
Baseline 27 MFCCs	16	65.6	70.9
DCTC/DCSC	16	67.5	73.0
BIG_REC Baseline 27 MFCCs	48	68.8	79.1
BIG_REC DCTC/DCSC	48	70.4	80.6

**Table 3.**  
27 feature TIMIT and RASC863 EVAL accuracies.

Parameter	Value
Frame Length	8 ms
Frame Spacing	1 ms
Frequency Warping $g(f)$	Bilinear, $\alpha = 0.4$
Number of DCTCs	15
Number of DCSCs	5
Frames per Block	251 ms (251frames)
Block Spacing	7 ms (7 frames)
Time Warping ( $dh/dt$ term)	Kaiser window, $\beta = 40$

**Table 4.**  
Optimum parameter settings for large feature set.

typically used filterbank front ends can be viewed as a unified framework. The reported experimental results can be explained using the unified time-frequency basis vectors as a common yardstick. First, Experiments A1 and A2 show that for static features, the proposed continuous Mel-shape warping results in slightly better performance than that obtained using Mel filterbank-derived basis vectors. By comparing their unified static basis vectors in **Figure 7(a)** and **Figure 10(a)**, our conjecture is that the quantization effect of the filterbank caused this difference. However, since the continuous Mel-shape warping and the filterbank are essentially two ways of implementing a Mel warping, the difference should be small as verified by the experimental results. It should be pointed out that it was experimentally verified that the standard way of implementing the MFFC front end, and MFFCs computed using unified basis vectors, result in identical feature values, provided the amplitude nonlinearity immediately follows the spectral magnitude step. Similarly, by comparing the unified dynamic basis vectors in **Figure 7(b)** and **Figure 10(b)**, it's clear that

TIMIT Database	Number of HMM mixtures	EVAL Accuracy (%)	Training Accuracy (%)
Baseline 39 MFCCs	32	69.7	76.2
DCTC/DCSC	32	72.5	79.4
BIG_REC Baseline 39 MFCCs	96	71.0	84.5
BIG_REC DCTC/DCSC	96	74.0	87.1
RASC863 Database	Number of HMM mixtures	EVAL Accuracy (%)	Training Accuracy (%)
Baseline 39 MFCCs	32	71.5	80.9
DCTC/DCSC	32	73.3	83.8
BIG_REC Baseline 39 MFCCs	64	72.0	85.0
BIG_REC DCTC/DCSC	64	74.2	87.1

**Table 5.**  
 75 feature TIMIT and RASC863 EVAL accuracies.

the non-uniform time resolution for a long segment of speech is a better representation of the spectral trajectory than the discrete time derivatives (most obvious in the zeroth order unified basis vectors). The more significant improvements over the baseline MFCC for various numbers of features in Experiments B and C support this observation.

Another set of experiments was conducted to address the issue of statistical significance. The goal was to show that the difference or similarity between the reported best cases for the DCTC/DCSC front end and the best baseline results in each previous experiment were statistically significant rather than due to noise or other random factors. These significance tests were conducted using the TIMIT database. To do this, the best results of the proposed method and the baseline were viewed as two random variables whose mean values were denoted as  $\mu_T$  and  $\mu_B$ . Then the 672 utterances of the TIMIT DEV and TIMIT EVAL sets were divided into 12 groups respectively, and test results were obtained for each group as samples. Since it's reasonable to assume the same (but unknown) variance for the proposed front end and the baseline

Experiment number	Hypothesis tested	Results
Exp. A1/A2, DEV set	$\mu_T > \mu_B$	Significant at 90% confidence level
Exp. A3, DEV set	$\mu_T = \mu_B$ ( $\mu_T$ uses a warping of 0.45)	Significant at 97.5% confidence level
Exp. B1/B2, DEV set	$\mu_T - \mu_B \geq 1\%$	Significant at 90% confidence level
Exp. B3, DEV set	$\mu_T - \mu_B \geq 2.5\%$	Significant at 97.5% confidence level
Exp. C1 (both 16 and 80 mixtures, EVAL set)	$\mu_T - \mu_B \geq 2\%$	Significant at 90% confidence level
Exp. C2 (both 32 and 96 mixtures, EVAL set)	$\mu_T - \mu_B \geq 2.5\%$	Significant at 97.5% confidence level

**Table 6.**  
 Results of statistical significance tests for reported TIMIT experiments.

(because the database was identical in all cases), a  $t$ -test with 22 degrees of freedom was performed to test the significance of the difference term, i.e.  $\mu_T - \mu_B$ . The results of these tests are summarized in **Table 6**.

#### 4.5 Frequency-dependent time warping in DCSC/DCTC scheme

In addition to the DCTC/DCSC implementation in which the time warping is independent of frequency, a slate of experiments for the DCSC/DCTC variation, which incorporated frequency-dependent time warping, was also conducted. The goal was to test the effectiveness of the auditory time-frequency trade-off caused by the nonlinear frequency selectivity for improving ASR performance. Specifically, the best warping factors obtained in the DCTC/DCSC experiments (i.e. 50 in the 27 feature case and 40 in the 75 feature case) were used as a baseline; smaller time warping for lower frequencies and larger time warping for higher frequencies were used with averages fixed at the baseline values (the block length was identical for all frequencies). Another equivalent method implemented was to use a longer block length for low frequencies compared to higher frequencies with the warping factor fixed. The results of these experiments showed no advantages over the baseline, which uses uniform time warping over all frequencies. This seems to imply that despite the results from human auditory research, which shows that humans have frequency-dependent temporal sensitivity [34–36], it may not play a crucial role, at least for the phone recognition ASR task evaluated in this chapter. Similar findings were observed by others. In one detailed study using wavelet signal processing to extract features for phonetic class recognition [51], the best performance obtained with wavelet features was only comparable to that obtained with MFCC features. In another study [52], a set of spectral-temporal features, which also accounts for the similar time-frequency trade-off, resulted in improved performance but only for restricted tasks (an isolated phone classification task rather than a continuous recognition application). The method introduced in [52] has not been adopted by the ASR community for general use.

### 5. Conclusion and future work

This chapter presents a generalized spectral-temporal feature extraction front end for representing speech information. The feature set is motivated by the attempt to mimic two primary properties of human hearing: frequency and time resolution. Based on a set of frequency and time-warping functions built into a set of modified 2-D cosine basis vectors and the trade-off between frequency and temporal and time resolution can be explored. A wide range of ASR experiments were conducted using the DCTC/DCSC method to comprehensively evaluate spectral-temporal resolution effects. This was done by adjusting parameters controlling the DCTC and DCSC parameters emphasize either spectral resolution or temporal resolution, and attempting to find the best overall “balance” point. The best combination point, using phonetic recognition experiments with the English language, also worked well with the Mandarin language.

Empowered by the front end unification approach, a higher level systematic unification can be envisioned. Conceptually, a recognizer front end should only require static features, with temporal patterns modeled by the recognizer. The human auditory system primarily performs spectral analysis whereas higher levels of



processing in the human brain appear to extract the longer terms spectral-temporal information. Apparently, the HMM framework is not able to adequately capture the temporal patterns contained in sequences of static speech features alone. Thus, it is possible that modeling of the “hidden” spectral-temporal patterns can be exploited by a data-driven training of a state-of-the-art recognizer, such as a deep neural network (DNN), which has the power of performing “deep learning.”

## **Author details**

Stephen A. Zahorian<sup>1\*</sup>, Xiaoyu Liu<sup>2</sup> and Roozbeh Sadeghian<sup>3</sup>

1 Binghamton University, Binghamton, USA


2 Dolby Laboratories Inc., San Francisco, USA

3 Harrisburg University of Science and Technology, Harrisburg, USA

\*Address all correspondence to: [stephen.zahorian16@gmail.com](mailto:stephen.zahorian16@gmail.com)

## **IntechOpen**

---

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Zahorian SA. Detailed Phonetic Labeling of Multi-Language Database for Spoken Language Processing Applications. Rome, NY, USA: Air Force Research Laboratory Information Directorate; 2015. Available from: <http://www.oracle.com/us/corporate/citizenship/corporate-citizenship-report-2563684.pdf>. DOI: 10.21236/ada614725
- [2] Peterson GE, Barney HL. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*. 1952;**24**(2):175-184. DOI: 10.1121/1.1906875
- [3] Hermansky H. Perceptual linear prediction analysis of speech. *The Journal of the Acoustical Society of America*. 1990;**87**(4):1738-1752. DOI: 10.1121/1.399423
- [4] Weber K, Wet F, Cranen B, Bodes L, Bengio S, Bourlard H. Evaluation of formant-like features for ASR. *Int. Conf. on Spoken Language (ICSLP)*. 2002. DOI: 10.1121/1.1781620
- [5] Garner P, Holmes W. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. *Proceedings of ICASSP*. 1998:1-4. DOI: 10.1109/ICASSP.1998.674352
- [6] Holmes J, Holmes W, Garner P. Using formant frequencies in speech recognition. *Proceedings of EUROSPEECH'97*. 1997;**4**:2083-2086
- [7] Bogert BP, Healy MJR, Tukey JW. The quefrequency analysis of time series for echoes: Cepstrum, pseudo autocovariance, cross-cepstrum and Saphé cracking. In: Rosenblatt M, editor. Chapter 15. *Proceedings of the Symposium on Time Series Analysis*. New York: Wiley; 1963. pp. 209-2243
- [8] Zwicker E, Fastl H. Chapter 3. In: *Psychoacoustics, Facts and Models*. Springer-Verlag; 1990. pp. 25-28
- [9] Stevens SS, Volkman J, Newman EB. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*. 1937;**8**(3):185-190. DOI: 10.1121/1.1915893
- [10] Fletcher H. Auditory patterns. *Reviews of Modern Physics*. 1940:12
- [11] Zwicker E. Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*. 1961;**33**(2):248-248. DOI: 10.1121/1.1908630
- [12] Glasberg BR, Moore BCJ. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*. 1990;**47**(1-2):103-138. DOI: 10.1016/0378-5955(90)90170-T
- [13] Bridle JS, Brown MD. An Experimental Automatic Word-Recognition System. JSRU Report. Vol. 1003. Ruislip, England: Joint Speech Research Unit; 1974
- [14] Patterson PD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand MH. Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K, editors. *Auditory and Perception*. Oxford, UK: Pergamon Press; 1992. pp. 429-446
- [15] Slaney M. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Apple Technical Report. Cupertino, CA: Advanced Technology Group, Apple Computer, Inc.; 1993. p. 35

- [16] Zhang X, Heinz MG, Bruce IC, Carney LH. A phenomenological model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*. 2001; **109**(2):648-670. DOI: 10.1121/1.1336503
- [17] Robinson DW, Dadson RS. A redetermination of the equal-loudness relations for pure tones. *British Journal of Applied Physiology*. 1956;**7**:166-181
- [18] Makhoul J. Linear prediction: A tutorial review. *Proceedings of the IEEE*. 1975;**63**:561-580. DOI: 10.1109/PROC.1975.9792
- [19] Memon S, Lech M, Maddage N. Speaker verification based on different vector quantization techniques with Gaussian mixture models. In: *Third Int. Conf. On Network and System Security*. 2009. pp. 403-408. DOI: 10.1109/NSS.2009.19
- [20] Jayanna HS, Prasanna SRM. Fuzzy vector quantization for speaker recognition under limited data conditions. *TENCON 2008-IEEE Region 10 Conference*. 2008:1-4. DOI: 10.1109/TENCON.2008.4766453
- [21] Chen J, Paliwal KK, Mizumachi M, Nakamura S. Robust MFCCs Derived from Different Power Spectrum. *Scandinavia: Eurospeech*; 2001
- [22] Wang C, Miao Z, Meng X. Differential MFCC and vector quantization used for real-time speaker recognition system. *IEEE Congress on Image and Signal Processing*. 2008: 319-323. DOI: 10.1109/CISP.2008.492
- [23] Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*. 1994;**95**(5):2670-2680. DOI: 10.1121/1.409836
- [24] Athineos M, Hermansky H, Ellis DPW. LPTRAPS: Linear predictive temporal patterns. In: *Proc. of Interspeech*. Jeju Island, Korea; 2004. pp. 1154-1157
- [25] Valente F, Hermansky H. Hierarchical and parallel processing of modulation spectrum for ASR applications. *ICASSP*. 2008:4165-4168. DOI: 10.1109/ICASSP.2008.4518572
- [26] Kleinschmidt M. Methods for capturing spectro-temporal modulations in automatic speech recognition. *Acustica united with acta Acustica*. 2002;**88**:416-422
- [27] Kleinshmidt M. *Localized Spectro-Temporal Features for Automatic Speech Recognition*. Switzerland: Eurospeech; 2003
- [28] Allen J. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, and Signal Processing*. 1977; **ASSP-25**(3):235-238. DOI: 10.1109/TASSP.1977.1163007
- [29] Kim C, Stern RM. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. *INTERSPEECH*. 2009:28-31. DOI: 10.21437/Interspeech.2009-5
- [30] Rao KR, Yip P. *Discrete Cosine Transform: Algorithms. Academic Press; 1990. Advantages. Applications*
- [31] O'Shaughnessy D. *Speech Communication: Human and Machine*. Addison-Wesley; 1987. p. 150
- [32] Smith JO, Abel JS. The bark bilinear transform. In: *Proceedings of the IEEE Workshop on Applications of Signal*

- Processing to Audio and Acoustics. New York; 1995. DOI: 10.1109/ASPAA.1995.482991
- [33] Wang S, Sekey A, Gersho A. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*. 1992;**10**(5):819-829. DOI: 10.1109/49.138987
- [34] Duifhuis H. Consequences of peripheral filter selectivity for nonsimultaneous masking. *The Journal of the Acoustical Society of America*. 1973;**54**(6):1471-1488
- [35] Bidelman GM, Khaja AS. Spectrotemporal resolution tradeoff in auditory processing as revealed by human auditory brainstem responses and psychophysical indices. *Neuroscience Letters*. 2014;**572**:53-57
- [36] Shailer MJ, Moore BCJ. Gap detection as a function of frequency, bandwidth, and level. *The Journal of the Acoustical Society of America*. 1983; **74**(2):467-473. DOI: 10.1121/1.389812
- [37] Meyer B, Ravuri SV, Schadler MR, Morgan N. Comparing different flavors of spectro-temporal features for ASR. *INTERSPEECH*. 2011:1269-1272. DOI: 10.21437/Interspeech.2011-103
- [38] Depireux DA, Simon JZ, Klein DJ, Shamma SA. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology*. 2001;**85**: 1220-1234. DOI: 10.1152/jn.2001.85.3.1220
- [39] Ge W. Two Modified Methods of Feature Extraction for Automatic Speech Recognition (Thesis). Binghamton: Department of Electrical and Computer Engineering, Binghamton University; 2013
- [40] Hermansky H, Morgan N. RASTA processing of speech. *IEEE Trans. Speech and Audio Processing*. 1994;**2**(4): 578-589. DOI: 10.1109/89.326616
- [41] Hermansky H, Sharma S. TRAPS - classifiers of temporal patterns. *ICSLP*. 1998;**3**:1003-1006
- [42] Liu X, Zahorian SA. A Unified Framework for Filterbank and Time-Frequency Basis Vectors in ASR Front Ends. Australia: ICASSP; 2015. DOI: 10.1109/ICASSP.2015.7178854
- [43] Chiu BY, Bhiksha R, Stern RM. Towards fusion of feature extraction and acoustic model training: A top down process for speech recognition. *INTERSPEECH*. 2009:32-35. DOI: 10.21437/Interspeech.2009-6
- [44] Chiu BY, Stern RM. Analysis of physiologically-motivated signal processing for robust speech recognition. *INTERSPEECH*. 2008:1000-1003. DOI: 10.21437/Interspeech.2008-291
- [45] Young S et al. The HTK Book (for HTK Version 3.4) Available from: <http://htk.eng.cam.ac.uk/>. Cambridge University; 2009 Revised for HTK Version 3.4
- [46] Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond. *Speech Communication*. 1990;**9**:351-356. DOI: 10.1016/0167-6393(90)90010-7
- [47] Lee K, Hon H. Speaker-independent phone recognition using Hidden Markov Models. *IEEE Trans. on Acoust., Speech, and Signal Processing*. 1989;**37**(11): 1642-1648. DOI: 10.1109/29.46546
- [48] Li A, Yin Z, Wang T, Fang Q, Hu F. RASC863-a Chinese speech corpus with four regional accents. Report of Chinese Academy of Sciences. 2004

[49] Nossair ZB, Silsbee PL, Zahorian SA. Signal Modeling Enhancement for Automatic Speech Recognition. Vol. 1. Proceedings of ICASSP; 1995. pp. 824-827. DOI: 10.1109/ICASSP.1995.479821

[50] Zahorian SA, Wong B. Spectral amplitude nonlinearities for improved noise robustness of spectral features for use in automatic speech recognition. The Journal of the Acoustical Society of America. 2011;**130**(4):2524. DOI: 10.1121/1.3655077

[51] Van Pham T. Wavelet analysis for robust speech processing and applications (thesis). Graz University of Technology. 2007

[52] Droppo JG III. Time-frequency features for speech recognition (thesis). University of Washington. 2000