NOVEL COCHLEAR FILTER BASED CEPSTRAL COEFFICIENTS FOR CLASSIFICATION OF UNVOICED FRICATIVES

Namrata Singh[1] , Nikhil Bhendawade[2] , Hemant A. Patil[3]

[1]Software Engineer, LG Soft India pvt. ltd., Embassy Tech Square, Bangalore, 560103, India.
[2] Design Engineer, Redpine signals pvt.ltd., Hitech City, Hyderabad - 500081, India.
[3]Dhirubhai Ambani Institute of Information Technology (DA-IICT), Gandhinagar-382007,India.

ABSTRACT

In this paper, the use of new auditory-based features derived from cochlear filters, have been proposed for classification of unvoiced fricatives. Classification attempts have been made to classify sibilant (i.e., /s/, /sh/) vs. non-sibilants (i.e., /f/, /th/) as well as for fricatives within each sub-category (i.e., intra-sibilants and intra-non-sibilants). Our experimental results indicate that proposed feature set, viz., Cochlear Filter-based Cepstral Coefficients (CFCC) performs better for individual fricative classification (i.e., a jump of 3.41 % in average classification accuracy and a fall of 6.59 % in EER) in clean conditions than the state-of-the-art feature set, viz., Mel Frequency Cepstral Coefficients (MFCC). Furthermore, under signal degradation conditions (i.e., by additive white noise) classification accuracy using proposed feature set drops much slowly (i.e., from 86.73 % in clean conditions to 77.46 % at SNR of 5 dB) than by using MFCC (i.e., from 82.18 % in clean conditions to 46.93 % at SNR of 5 dB).

KEYWORDS

Unvoiced fricative sound, auditory transform, cochlear filter cepstral coefficients, Mel cepstrum, sibilants, non-sibilants.

1. INTRODUCTION

Classification of appropriate short regions of speech signal into different phoneme classes (*e.g.,* fricatives *vs.* plosives) based on its acoustic characteristics is an interesting and challenging research problem. In this paper, we present an effective feature set for classification of one particular class of phonemes, *viz.*, *unvoiced* fricatives. Fricative sounds are very unique class of phonemes in the sense that for fricatives, the sound source occurs at the point of constriction in the vocal tract rather than at the glottis. There are two types of fricatives, *viz.*, voiced and unvoiced (having different speech production mechanisms). For example, in case of *voiced* fricatives, noisy characteristics caused by the constriction in the vocal tract are accompanied by vibrations of vocal folds, thereby imparting some periodicity into the produced sound. However, during the production of *unvoiced* fricatives, vocal folds are relaxed and not vibrating. This lack

DOI : 10.5121/ijnlc.2014.3402

of periodicity results in relatively more random waveform pattern. Furthermore, voiceless fricatives being noise-like, having highly *turbulent* source, are dynamic, relatively short and weak (*i.e.,* having low energy) making classification even more difficult, especially, due to severe *masking* of fricative sounds by noise (i.e., under signal degradation conditions.

Since production of unvoiced fricatives is governed by source (*e.g.,* frication noise originating from constriction in vocal tract) - filter (*i.e.,* oral cavity) model theory [1, 2], they may be distinguished depending on location of constriction in oral cavity. This constriction at different locations accounts for distinct acoustical characteristics. To reliably predict the characteristics of fricative sounds, two approaches could be considered, v*iz.,* modeling the production mechanism of fricatives [3] or modeling response of human ear corresponding to the acoustical characteristics of each fricative class [4,5]. Our study focuses on second approach. To that effect, we propose use of cochlear filters to model response of human ear. Since among the various acoustic cues (*e.g.,* amplitude, spectral, durational and transitional characteristics) used previously, spectral cues were found to be the most efficient; we have used spectral information as a basis of classification of *four* unvoiced fricatives, *viz.*, /f/, /th/, /s/ and /sh/.

The rest of the paper is organized as follows: Section 2 gives brief discussion of relevant literature that deals with the earlier attempts to classify fricative sounds using various *acoustic* features. Section 3 discusses the details of proposed feature set and gives the comparison between Fourier transform and auditory transform and its significance for unvoiced fricative classification**.** Section 4 describes the experimental setup which is followed by the comparison of classification results using proposed and baseline features under various experimental evaluation factors (*e.g.,* cross-validation, dimension of feature vector, number of sub-band filters and signal degradation conditions) in Section 5. Finally, Section 6 concludes the paper along with future research directions.

2. LITERATURE REVIEW

The earlier studies in the area of fricative sound classification used *Root Mean Square (RMS) amplitude* of fricative sound as an acoustic cue to distinguish between sibilants and non-sibilants [7, 8].

Study reported in [9] used *duration* of fricative noise as a perceptual cue to distinguish between sibilants and non-sibilants as they found that sibilants are on an average *33 ms* longer than non-sibilants. However, the approach had several issues such as *durational* features often vary with speaking rate and contextual complexity. In an different experiment, it was also found that listeners identify fricative sound using only the initial fraction of utterance contrary to the earlier conclusion reported in [9] that absolute fricative noise duration can be used as a *perceptual* cue [10]. Instead *relative duration* (*i.e.,* duration of fricative relative to entire word duration) was proposed in further studies [11]. This study found significant difference among all the places of articulation for fricative using *relative duration* as a cue, however, with the exception of unvoiced non-sibilants.

Various *spectral* features have been investigated and used for a long time since the hypothesis presented in [12], that spectrum of fricatives is governed by *size* and *shape* of resonance chamber in front of constriction point. Work presented in [13] supported this finding when the spectral characteristics of front (near-flat spectrum), middle (spectral peak around *3.5* kHz) and back (spectral peak around *1.5* kHz) unvoiced fricatives were examined. Though the locations of spectral peaks are influenced by speaker differences [14] and age differences among speakers [15], it was consistently observed in many studies that the spectral peaks of sibilants always lie between *1-6* kHz range while non-sibilants show almost *flat* spectrum extending beyond *8* kHz.

Previous studies depict that various acoustic cues have been found effective for distinguishing between sibilant and non-sibilant class as a whole and between fricatives within sibilant class. However, analyzing the characteristics of fricatives within non-sibilant class has proved less conclusive resulting in poor classification accuracy. In this paper, we propose an auditory-based approach, for relatively better analysis and distinction of non-sibilant sounds in both clean and noisy environments by using cochlear filters (which resemble impulse response of human cochlea to any sound event). As human ear could distinguish between fricative sounds better than any other classification system (both in clean and noisy conditions), spectral cues derived from application of cochlear filters have been used for distinction between all four unvoiced fricatives (*i.e.,* /f/, /th/, /s/ and /sh/). Results have also been reported for classification of sibilant *vs.* non-sibilant sounds and for fricatives within each subcategory (i.e., /f/ *vs.* /th/ and /s/ *vs.* /sh/).

3. COCHLEAR FILTER-BASED CEPSTRAL COEFFICIENTS (CFCC)

CFCC features (derived from auditory transform) have been proposed first time in [4] for speaker recognition application. Auditory transform is basically a *wavelet* transform, however, the mother wavelet (*i.e.,* $\psi(t)$) is chosen in such a manner that the cochlear filters (whose impulse response corresponds to dilated version of mother wavelet) emulate the cochlear filters present in cochlea of human ear. Cochlear filters are responsible for perception of sound by human auditory system and would thus be expected to include properties of *robustness* under noisy or signal degradation conditions (*i.e.,* may be better than most of the other artificial speech recognition or classification systems in noisy environments). The auditory transform is implemented as a bank of sub-band filters where each sub-band filter corresponds to the cochlear filter present along the *basilar membrane* (BM) in cochlea of human ear. These cochlear filters have been found to have a bandwidth that varies with their central frequencies. In particular, the bandwidth of these filters increases with increasing *central* frequency (*i.e.,* f_c) and has almost *constant quality factor* (*i.e.,* Q). These filters thus provide a range of analysis window durations and bandwidth for analyzing speech signal so that rapidly varying signal components are analyzed with shorter window duration than slowly varying components preserving the time-frequency resolution in both cases. Fig. 1 shows block diagram for implementation of CFCC [4, 5].

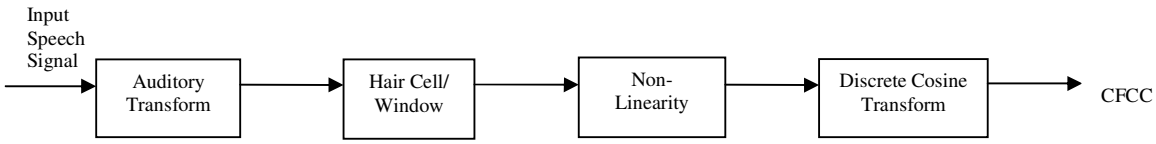


Fig. 1. Auditory-based feature extraction technique, *viz.,* Cochlear Filter based Cepstral Coefficients (CFCC) [4,5].

We have chosen *logarithmic* nonlinearity instead of *cubic root* nonlinearity used in earlier studies [4,5] as it resulted in better classification, *i.e.,*

$$y(i, j) = \ln(S(i, j)). \quad (1)$$

where S(i,j) is the nerve spike density, obtained from hair cell output for each sub-band with duration for nerve density count taken as 12 ms (i.e., =12 ms), calculated with window shift duration of 5ms.

3.1. Details of cochlear filters

Fig. 2 shows the frequency response of cochlear filters used in proposed feature set. Filters have been placed according to Mel scale and central frequencies of filters are calculated according to,

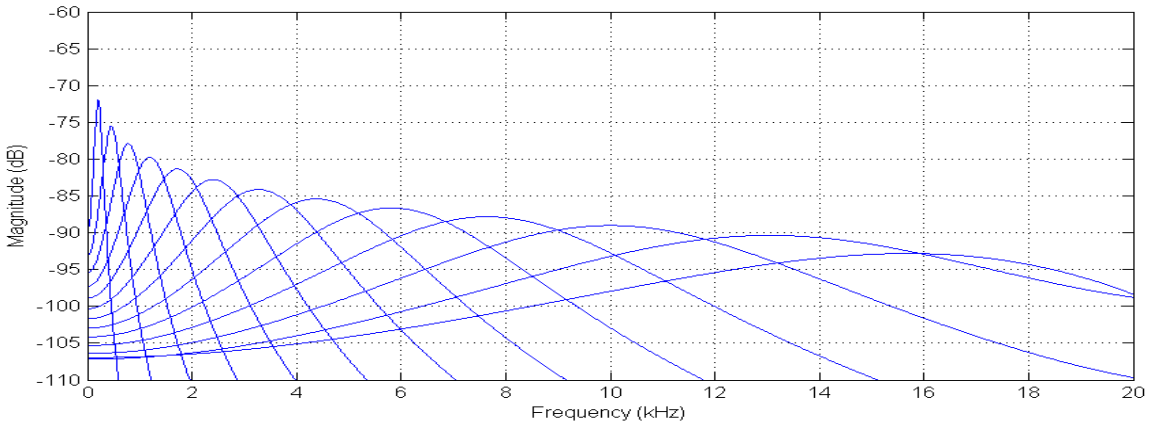


Fig. 2. Frequency response of *13* cochlear filters placed along Mel-scale with α=*2* and β=*0.45* .

$$f_{mel} = 1127 \ln(1 + \frac{f_{lin}}{700}), \quad (2)$$

where f_{mel} is central frequency along *Mel* scale and f_{lin} is corresponding central frequency along *linear* scale (*i.e.,* in Hz). Filters are placed uniformly along Mel scale so the distribution appears exponential along linear scale. Parameters α and β which decide filter shape have been optimized as *2* and *0.45*, respectively, (for database used in present work).

Though *13* cochlear filters have been used in our work (for the reasons described in Section 5.3), we experimented with number of sub-band filters to find the minimum number of cochlear filters required to capture the distinctive spectral characteristics of the unvoiced fricatives. Six filters have been found to be significant in our analysis (giving classification accuracy of *84.07 %*). Fig. 3 shows the frequency responses of these significant filters. Corresponding impulse responses have been shown in Fig. 4. It is noted that as central frequency of filters increases, bandwidth also increases maintaining a *near-constant Q* factor of *2.15* (as shown in Table 1). Furthermore, higher frequency components are analyzed with larger time resolution (shorter analysis window durations) while higher frequency resolution is used for analyzing lower frequency components. As shown in Fig. 4, frequency components near *13.1* kHz are analyzed with window of approximately 0.561 *ms* duration[1] while window of approximately *11.4 ms* is used for analyzing frequency components near *451* Hz. This is known as constant Q-filtering and this is what happens in Basilar membrane of human cochlea during speech perception

[1] Only half part of the analysis window has been displaced in Fig.4, since the window is *symmetric.*

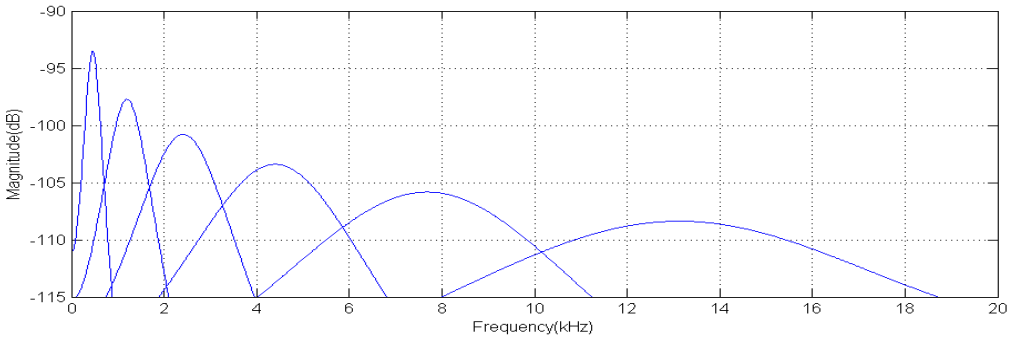


Fig. 3. Frequency response of six cochlear filters found significant for unvoiced fricative classification.

Table 1 : Central frequencies of cochlear filters found significant for unvoiced fricative classification

Cochlear filter Index	**Center frequency(Hz)**	***-3dB* Bandwidth(B) (Hz)**	**Quality factor** (f_c/B)
1	451	210	2.1476
2	1191	550	2.1654
3	2408	1120	2.1500
4	4408	2050	2.1502
5	7696	3580	2.1497
6	13100	6450	2.04

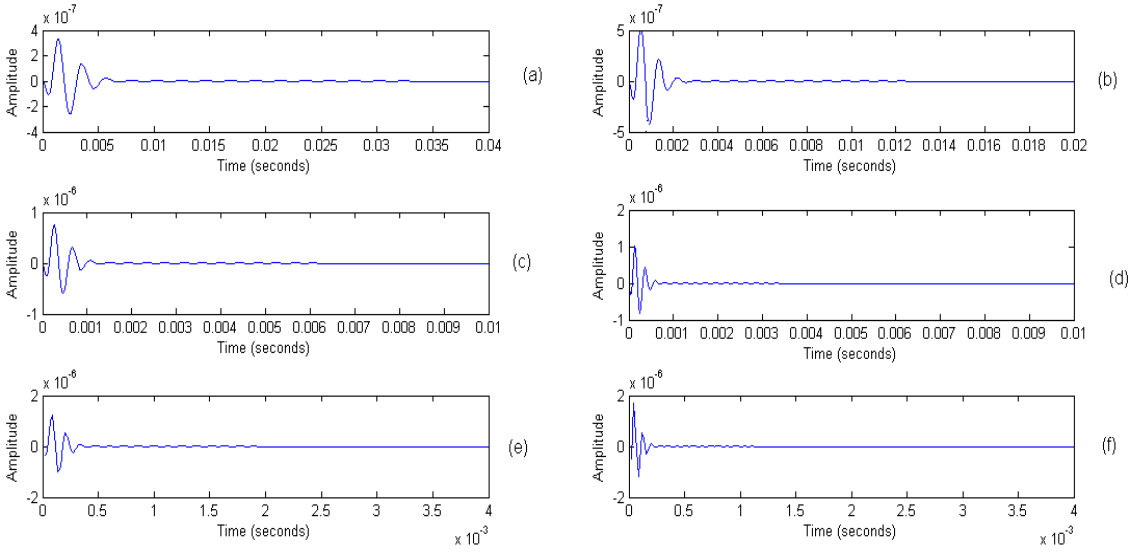


Fig.4. Impulse response of six cochlear filters found significant for unvoiced fricative classification with central frequencies (a) *451* Hz, (b) *1191* Hz, (c) *2408* Hz, (d) *4408* Hz, (e) *7696* Hz, and (f) *13.1* kHz.

3.2. Short-time Fourier transform *vs.* Auditory transform:

Short-time Fourier transform (STFT) is the most widely used technique for analyzing the frequency-domain characteristics of localized regions of speech signal. Though efficient, it uses *fixed length* window for signal analysis resulting in *constant* time-frequency resolution and hence *improving* resolution in time-domain will result in *degradation* of resolution frequency-domain (*i.e.,* Heisenberg's uncertainity principle in signal processing framework [16]). In addition, several optimized algorithms used in evaluating STFT *via* Fast Fourier Transform (FFT), add to the computational noise, by increasing computational speed at the expense of slight compromise in accuracy. This might seriously affect spectral cues in case of non-sibilants as they have weak resonances (*i.e.,* formants) in their spectrum. Fig. 5-Fig.8 gives the comparison between the spectrum derived from auditory transform and traditional Fourier transform. Each spectrum is averaged from initial, middle and end regions of fricative sounds for each fricative class such that it represents the overall average spectral characteristics for that class. Hamming window with window duration of *12 ms* with frame rate of *5 ms* has been used for FFT-based computation of Fourier transform while auditory transform is computed using *13* cochlear filters of variable length by the procedure described in [17]. Fourier transform spectrum is affected by regular spikes because of the fixed window duration for all frequency bands (as seen in the Fig. 5 in the form of periodic spikes in spectrum, as shown in Fig. 5-Fig.8). On the other hand, spectrogram generated from auditory transform provides flexible time-scale resolution by employing variable length filters and hence it is free from these spikes and also preserves information about formant frequencies [4, 5]. From Fig. 7 and FIg. 8, it is also clear that sibilants show spectral peaks around *5* kHz while such energy concentration at particular frequency is absent in non-sibilants and they tend to have near-flat spectrum (which is shown Fig. 5 and Fig.6). The reason for this could be explained from speech production mechanism. In particular, during production of sibilant sounds, point of constriction lies near alveolar ridge resulting in considerable length of front cavity, (created between point of constriction and lips) which in turn is responsible for spectral filtering of the turbulant sound produced from the constriction introducing resonances into the spectrum while such spectral filtering is almost absent in case of *labiodental* (/f/) and *interdental* (/th/) non-sibilants as point of constriction itself lies at lips in the former case while between upper and lower teeth in later ([18], [19]).

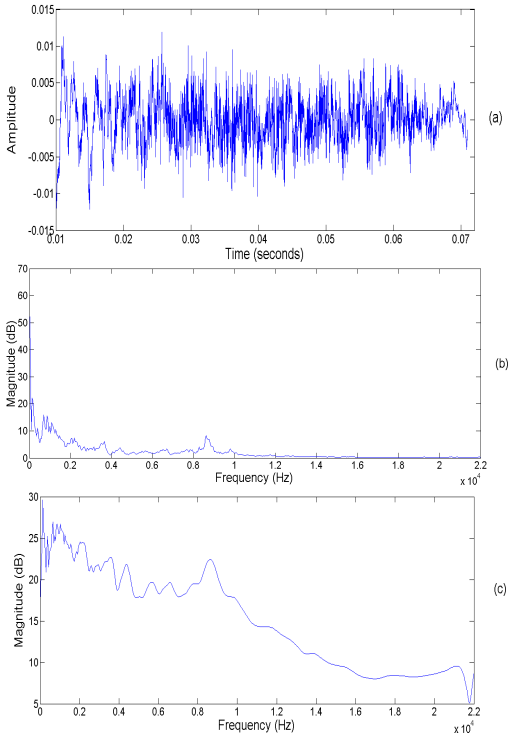


Fig.. 5 (a) Waveform for fricative sound /f/ (i.e., non-sibilant) and corresponding spectrum using (b) Fourier transform (c) auditory transform.

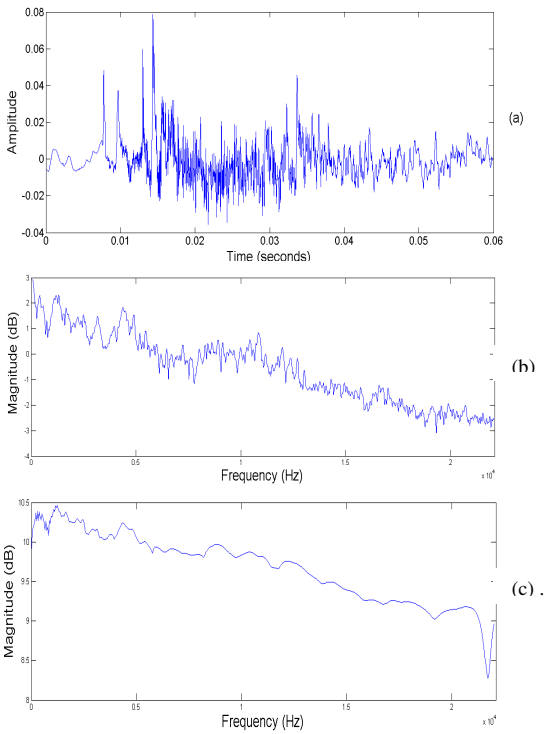


Fig.. 6 (a) Waveform for fricative sound /th/ / (i.e., non-sibilant) and corresponding spectrum using (b) Fourier transform (c) auditory transform..

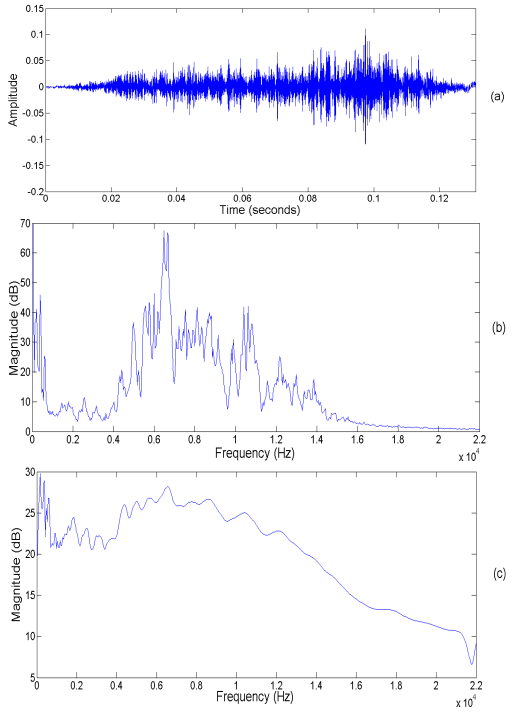


Fig.. 7 (a) Waveform for fricative sound /s/ (i.e., sibilant) and corresponding spectrum using (b) Fourier transform and (c) auditory transform..

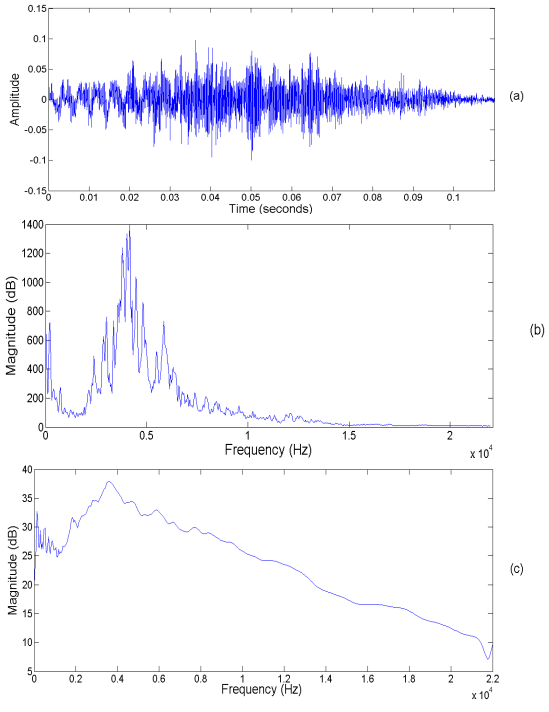


Fig. 8 (a) Waveform for fricative sound /sh/ and corresponding spectrum using (b) Fourier transform and (c) auditory transform..

3.

The Mel scale filterbank has triangular shaped sub-band filters which are *not* smooth at the vertex of each triangle [20]. On the other hand, from Fig. 3, it is evident that cochlear filters have bell-shaped frequency response and hence are relatively much more smoother than the Mel filters. This smoothness of the cochlear filters may help in suppressing the noise.

Robustness of CFCC features could also be explained from similarity of auditory transform with signal processing abstraction of cochlea in human ear. In noisy acoustic environment, human listeners perform robustly. In particular, human hearing system is robust to the noise because of *amplification* mechanism in auditory transform to take care of mechanical vibrations of eardrum at the threshold of hearing (*i.e.*, $2\times10^{-5} N/m^2$) [21]. To support this observation, study reported in [22] claims that two or more rows of outer hair cells (OHC) in the cochlea are pumping fluid which accelerates the process of detecting sub-band energies in speech sound. In addition, those OHC might be setting up their own *vortex* to act as the amplifier [21]. The sub-band-based processing and energy detection comes from the original studies reported in [23]. Study in [23] is based on belief that human ear is a *frequency analyzer*, except for detection of transient sounds. In this context, CFCC employs continuous-time wavelet transform (CWT) which has mother wavelet $\psi(t)$ to aid for noise suppression and to detect the *transitional* sounds such as fricatives. This is analyzed below.
we have eq. (3) from [4],

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0; \tag{3}$$

$$\int_{t=-\infty}^{+\infty} \psi(t)dt = 0 = \int_{t=-\infty}^{+\infty} t^0\psi(t)dt. \tag{4}$$

This means that $\psi(t)$ has *one* vanishing moment and it will suppress polynomial of degree *zero* [16]. Let $f(t)$ be the clean speech signal, $w(t)$ be the additive white noise signal, then the noisy speech signal, $x(t)$, is given by

$$x(t) = f(t) + w(t). \tag{5}$$

Taking wavelet transform on both sides and using *linearity* property of CWT, we get,

$$Wx(a,b) = Wf(a,b) + Ww(a,b), \tag{6}$$

where $Wf(a,b) = \int_{t=-\infty}^{+\infty} f(t)\psi_{a,b}^{*}(t)\,dt = \langle f, \psi_{a,b} \rangle = \int_{t=-\infty}^{+\infty} f(t)\frac{1}{\sqrt{a}}\psi^{*}\left(\frac{t-b}{a}\right)dt$ and the symbol $\langle , \rangle$ denotes the *inner product* operation and $Wf(a,b)$ means CWT of signal $f(t)$. Hence, eq. (6) becomes,

$$\therefore \langle x, \psi_{a,b} \rangle = \langle f, \psi_{a,b} \rangle + \langle w, \psi_{a,b} \rangle, \tag{7}$$

$$\therefore \left|\langle x, \psi_{a,b} \rangle\right| = \left|\langle f, \psi_{a,b} \rangle + \langle w, \psi_{a,b} \rangle\right|,$$

$$\Rightarrow \left|\langle x, \psi_{a,b} \rangle\right| \le \left|\langle f, \psi_{a,b} \rangle\right| + \left|\langle w, \psi_{a,b} \rangle\right|. \tag{8}$$

It is well known that the Taylor formula relates the differentiability of a signal $f(t)$ to local polynomial approximation. Let us assume that signal $w(t)$ is m times differentiable in $[v-h, v+h]$. If $P_v(t)$ is Taylor polynomial in the neighborhood of point v, then

$$w(t) = P_v(t) + \varepsilon_v(t), \tag{9}$$

where the approximation error $\varepsilon_v(t)$ is refined by non-integer exponent α (called as *Lipchitz exponent* or *Holder exponent* in mathematical literature) . In particular, there exists $K > 0$ such that

$$\forall t \in \mathbb{R}, \quad \left|\varepsilon_v(t)\right| = \left|w(t) - P_v(t)\right| = K\left|t - v\right|^{\alpha}. \tag{10}$$

Let mother wavelet $\psi(t)$ has n vanishing moments and signal $w(t) \in L^2(\mathbb{R})$ (i.e., Hilbert space of finite energy signals) has *non-integer* Lipchitz exponent α. In this case, we have following two theorems [chapter 6, pp. 169-171, 24], [25].

Theorem 1: If the signal $w(t) \in L^2(\mathbb{R})$ is uniformly Lipchitz $\alpha \le n$ over the closed interval $[b_1, b_2]$ then there exists $K > 0$ such that

$$\forall (a,b) \in \mathbb{R}^+ \times [b_1, b_2], \qquad \left|\langle w, \psi_{a,b} \rangle\right| \le K a^{\alpha + \frac{1}{2}}. \tag{11}$$

Theorem 2 (JAFFARD) : If the signal $w(t) \in L^2(\mathbb{R})$ is Lipchitz $\alpha \le n$ at a point v, then there exists $K > 0$ such that

$$\forall (a,b) \in \mathbb{R}^+ \times \mathbb{R}, \qquad \left|\langle w, \psi_{a,b} \rangle\right| \le K a^{\alpha + \frac{1}{2}} \left(1 + \left|\frac{b-v}{a}\right|^{\alpha}\right), \tag{12}$$

where a and b are *scale* and *translation* paramters in the definition of CWT. It should be noted that *converse* is also true for both the above theorems. Since in present case, from eq. (4), we have $n = 1$ which implies that $\alpha \le 1$. Above two theorems gives a guarrantee that the wavelet transform of noise signal will decay faster as the scale parameter goes to zero (i.e., the at the fine scales). On the other hand, for larger values of scale parameter, it does not introduce any constraint. In particular, due to Cauchy-Schwartz inequality, we have

$$\left|\langle w, \psi_{a,b} \rangle\right| \le \|w\| \quad \|\psi_{a,b}\| = \|w\| \quad \|\psi\|. \tag{13}$$

Since due to normalization of mother wavelet, $\|\psi\| = \|\psi_{a,b}\| = 1$ [chapter 4, 24], we have,

$$\left|\langle w, \psi_{a,b} \rangle\right| \le \|w\|. \tag{14}$$

Hence, the wavelet transform of noise signal is bounded by $\|w\|$, at larger scale parameter. From eq.(8) and eq. (11), we have,

$$\left|\langle x, \psi_{a,b} \rangle\right| \le K_1 a^{\alpha_1 + \frac{1}{2}} + K_2 a^{\alpha_2 + \frac{1}{2}}.$$

Similarly, eq.(8) and eq. (11), we have

$$\left|\left\langle x,\psi_{a,b}\right\rangle\right| \le K_1 a^{\alpha_1+\frac{1}{2}}\left(1+\left|\frac{b-v}{a}\right|^{\alpha_1}\right)+K_2 a^{\alpha_2+\frac{1}{2}}\left(1+\left|\frac{b-v}{a}\right|^{\alpha_2}\right), \quad (15)$$

where $K_1, K_2 > 0$ and α_1 and α_2 are the Lipchitz exponents of clean speech signal and additive white noise, respectively. Since, wavelet transform of noise signal will decay, it is evident from eq. (14) and eq. (15) that additive noise is *suppressed* in wavelet-domain. Since, CFCC inherently employs CWT representation to mimic cochlear filters in human hear, it is expected that CFCC will have noise suppression capability. This is also demonstrated with experimental results for unvoiced fricative classification under noisy conditions in Section 5.4.

4. EXPERIMENTAL SETUP

4.1.Database used in this study

Preparation of sufficient training and testing data for each fricative involves extracting fricatives sounds from continuous speech in different contexts (of speech recordings) from different speakers. All the fricatives have been manually extracted (using Audacity software [26]) from CHAINS database [27] of continuous speech in solo reading style (recorded using a Neumann U87 condenser microphone). The database is publicly available having *4* extracts (*viz.*, rainbow text, members of the body text, north wind text and Cinderella text), a set of *24* sentences having text material corresponding to TIMIT database and a set of other *9* CSLU's Speaker Identification Corpus sentences.

Table 2 summarizes the details (such as number of speakers and contexts of fricative sounds) of the dataset for each fricative sound used in this work. Words for segmenting fricative samples are collected such that samples consist of variety of contexts. Column 5 in Table 2 gives this contextual information (*i.e.,* underlined region in a word indicates the location of fricative sound).

Table 2. Training and testing data extraction for each unvoiced fricative class

		# of samples	**# of Speakers (Male+Female)**	**Context associated with training and testing samples**
Non-Sibilants	/f/	208	5 (1 M + 4 F)	for, of, affirmative, find, enough, fire, fish frantically, fortune, frightened, fairy, forgot, fifth, if, fires, off, beautiful, form, refused, few, fell, from, food, roof, centrifuge and Jeff
	/th/	143	21 (11 M + 10 F)	Thought, teeth, North, tooth, think, throughout, everything, path, something and mouth
Sibilants	/s/	305	6 (2 M + 4 F)	sun, sunlight, say, looks, support, must, necessary, receive, dance, sing, small, surface, cost, sloppy, appearance, same, atmosphere, escape, sermon, subdued, task, rescue, ask, suit, saw, system, loss and centrifuge
	/sh/	254	14 (4 M + 10 F)	wash ,under-wash ,discussion ,condition, shine, share, shape, shelter, shotguns, action, Trish and she
	Total	910	23 (10 M + 13 F)	

M-Male, F-Female

4.1. Front end analysis

To evaluate the relative performance of the proposed feature set, state-of-the-art feature set, *viz.,* MFCC is used as the *baseline* feature set. Front end analysis involves computation of both CFCC and MFCC features from corresponding spectra. Spectral analysis is done using Discrete Fourier Transform (DFT) up to *22.05* kHz (corresponding to sampling frequency of *44.1* kHz) as it was observed previously that spectral information of non-sibilants extend above *10* kHz [28]. Frame size of *12 ms* along with Hamming window and frame rate of *5 ms* is used for computation of MFCC features while CFCC features are computed as described in Section 3 . Though such small window size of *12 ms* reduces the resolution in frequency-domain in case of MFCC, we observed that *temporal* development of fricative sounds can be better modeled using larger number of feature vectors per fricative sound (*i.e.,* small window size) thereby increasing time resolution, especially for non-sibilant /th/ which has average duration as small as *71.86 ms* (computed over *143* samples used in this study). Cepstral Mean Subtraction (CMS) is performed after MFCC and CFCC computation to take care of variations in recording devices and transmission channels. Furthermore, use of CMS also resulted in considerable increase in % classification accuracy.

4.3 Hidden Markov Model (HMM)

In this work, HMM is used as a pattern classifier since it preserves the *temporal* development of the fricative utterance which is often important in perception of fricative sounds. On the other hand, temporal variation is irrelevant in other widely used techniques such as discriminatively-trained pattern classifier, *viz.,* support vector machines (SVMs) in which classification is done independently for each frame in an utterance [31]. HMM evaluates the probability of an utterance being particular fricative sound based on observation and transition probabilities of observed sequence . A *3*-state continuous density HMM has been employed for modeling of each fricative class.

4.4 Performance measures

To facilitate the performance comparison between proposed and baseline feature sets, three performance measures, *viz.,* classification accuracy, % Equal Error Rate (EER) and minimum Detection Cost Function (DCF) have been employed. % classification accuracy is defined as, %

$$\text{Classification Accuracy} = \frac{\#\text{ of test samples correctly identified } (N_c)}{\text{Total \# of test samples } (N_t)} \times 100. \qquad (16)$$

Error is a measure of misclassification probability. Classification error could be due to failure of a classifier to detect a true test sample or due to acceptance of false test sample. We have used Detection Error Trade-off (DET) curve for analyzing the error rates which gives the trade-offs between missed detection rate (*i.e.,* miss probability) and false acceptance rate (*i.e.,* false alarm probability) [32]. Two performance measures, *viz.*, % Equal Error Rate (EER) and minimum Detection Cost Function (DCF) have been employed for quantifying the error associated with classification task. % EER corresponds to an optimal classification threshold at which both the errors (*i.e.,* false acceptance and missed detection) are *equal* while DCF calculates the minimum cost associated with the errors by penalizing each error according to its relative significance. DCF is given by,

$$DCF = C_{\text{miss}} * P_{miss} * P_{true} + C_{fa} * P_{fa} * P_{false}, \qquad (17)$$

where P_{miss} and P_{fa} are *missed* detection and *false alarm* probabilities while C_{miss} and C_{fa} are costs associated with them. P_{true} and P_{false} denote prior probabilities of *true* and *false* samples, respectively, which in turn depends upon number of *genuine* and *imposter* trials performed. We have employed *equal* penalties to both the errors (*i.e.*, $C_{miss} = C_{fa} = 1$) for evaluating DCF. We have also reported *95 %* confidence intervals of classification accuracy to quote *statistical* significance of our experimental results. Confidence intervals have been estimated by parametric techniques [33].

5. EXPERIMENTAL RESULTS

In this section, experiments are performed to evaluate the proposed feature set for various experimental evaluation factors such as cross-validation, effect of feature dimension, number of sub-band filters and robustness against signal degradations. The details of these experiments and analysis of results are presented in next sub-sections.

5.1. Fricative Classification using CFCC and MFCC

Using *13*-dimensional feature vector (for both CFCC and MFCC feature sets),following three classification tasks are performed on *2*-fold cross-validated data.

1. Modeling sibilants and non-sibilants as different classes,
2. Modeling fricatives within sibilants and non-sibilants as different classes (*e.g.*,/s/ *vs.* /sh/ and /f/ *vs.* /th/),
3. Modeling each kind of fricative sound as a different class.

Table 3 shows the overall classification results for above classification tasks followed by individual class analysis depicted via confusion matrices (shown in Table 4 -Table11). Corresponding DET curves have been shown in Fig. 9, Fig. 10 and Fig. 11, respectively. Following observations could be made from the results.

a. CFCC features perform consistently superior to baseline feature set (*i.e.,* MFCC) in all three classification tasks as mentioned above(Table 3 to Table 11).
b. CFCC improves the overall % classification accuracy of sibilant *vs.* non-sibilant classification (*i.e.,92.01 %*, as shown in Table 3) by improving the rate of identifying genuine non-sibilant samples (*i.e.,90.15 %*, as shown in Table 5) while genuine sibilant samples have been identified equally well using both MFCC and CFCC feature sets (Table 4 and Table 5). DET curve (shown in Fig. 9) indicates that CFCC performs better than MFCC at *all* the operating points of the curve (*i.e.,* by varying *classification threshold*) reducing % EER by*6.37%*.
c. Classification within sibilant class is much more accurate than within non-sibilant class in case of both feature sets (*i.e.,* MFCC and CFCC). Furthermore, classification accuracy within sibilant class is almost same for both features, while % EER has been significantly reduced in case of CFCC (by *5.37 %*) suggesting that overlapping score distribution of genuine and imposter test samples in case of MFCC has been considerably reduced by using proposed CFCC (Table 3, Table 8 and Table 9, Fig. 10(b)).
d. Though classification accuracy within non-sibilant class has been improved in case of CFCC (because of better identification of genuine /th/ test samples), the % EER is much higher in case of both features (Table 3, Table 6 and Table 7, Fig. 10(a)).
e. Individual classification analysis of all fricatives also shows the effectiveness of proposed feature set to better identify genuine /th/ test samples than MFCC resulting in overall

superior performance (Table 3, Table 10 and Table11). DET curve (shown in Fig. 11) also depicts the superiority of CFCC which performs better than MFCC for *most* of the operating points of the DET curve (of varying classification threshold) reducing % EER by *6.59 %*.

Table 3: Comparison of classification results using CFCC and MFCC

	Average % classification accuracy		**% EER**		**Minimum DCF**	
Feature set → / Task ↓	MFCC	CFCC	MFCC	CFCC	MFCC	CFCC
Sibilants *vs.* Non-sibilants	89.44 [86.62, 92.26]	**92.01 [89.52, 94.5]**	27.91	**21.54**	0.2780	**0.2121**
/f/ *vs.* /th/ (*i.e.*, within non-sibilant class)	76.42 [70.15, 82.69]	**83.18 [77.66, 88.7]**	31.77	**25.52**	0.3151	**0.2782**
/s/ *vs.* /sh/ (*i.e.,*within sibilant class)	96.45 [94.28, 98.62]	**97.55 [95.74,99.36]**	21.14	**15.77**	0.13	**0.10**
All four fricatives (*i.e.,*/f/,/s/,/sh/,/th/)	85.73 [82.52,88.94]	**89.14 [86.29 92]**	26.37	**19.78**	0.2148	**0.1549**

To summarize, sibilants are classified accurately by using both feature sets, MFCC and CFCC. Interestingly, within non-sibilants, /f/ is classified equally well in both feature sets, however, classification accuracy of /th/ is much higher in case of CFCC as compared to the MFCC. The reason for this could be large spectral variation in /th/ sound. /f/ sound is found to occupy weak spectral resonances around *1.5* kHz and *8.5* kHz. However, such energy concentration is *not* observed consistently with all the /th/ test samples. On the other hand, spectral distribution of /th/ sound is highly variable (especially above *8* kHz) across different speakers and contexts. As CFCC incorporates cochlear filters and several processes involved in auditory perception of sound (*eg.,* neural firings, nerve spike density, etc.), the spectral variability in /th/ sound may bebetter modelled (as it happens in human auditory system) by CFCC resulting in considerable increase in classification accuracy of /th/ as compared to MFCC.

Table 4: Confusion matrix showing % classification accuracy for sibilant *vs.* non-sibilant classification using MFCC features

Identified → / Actual↓	Non-sibilants	Sibilants
Non-sibilants	83.68	16.32
Sibilants	6.96	93.05

Table 5: Confusion matrix showing % classification accuracy for sibilant *vs.* non-sibilant classification using CFCC features

Identified / Actual↓	Non-Sibilants	Sibilants
Non-sibilants	**90.15**	9.85
Sibilants	6.82	**93.18**

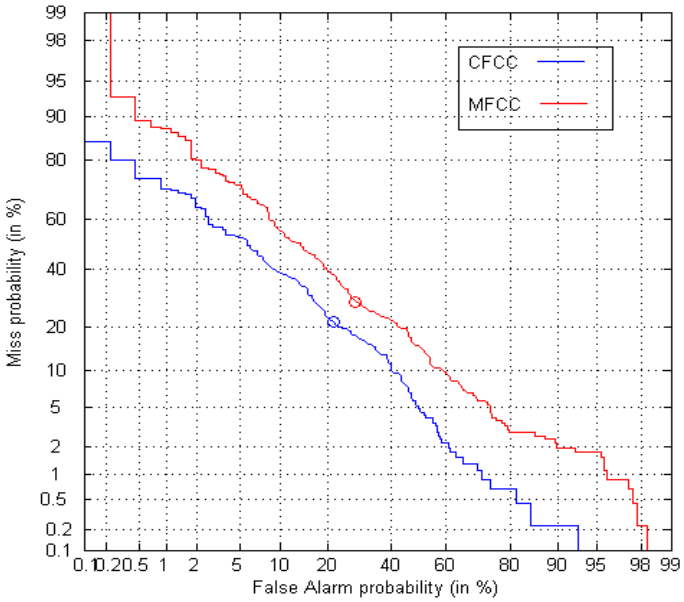


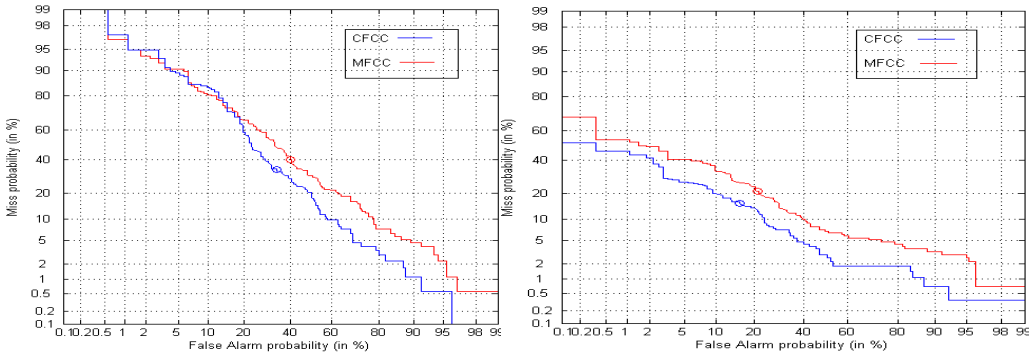
Fig.9. DET curves for sibilant *vs.* non-sibilant classification using baseline and proposed feature sets.

Table 6: Confusion matrix showing % classification accuracy of classification within non-sibilants using MFCC

Identified → / Actual ↓	/f/	/th/
/f/	85.34	14.66
/th/	36.46	63.54

Table 7: Confusion matrix showing % classification accuracy of classification within non-sibilants using CFCC

Identified → / Actual ↓	/f/	/th/
/f/	**86.11**	13.89
/th/	21.04	**78.96**



(a) (b)

Fig.10. DET curves for classification using baseline (MFCC) and proposed (CFCC) feature sets (a) within non-sibilant class (*i.e.*, /f/ *vs.* /th/) (b) within sibilant class (*i.e.*, /s/ *vs.* /sh/).

Table 8: Confusion matrix showing % classification accuracy of classification

Identified → / Actual ↓	/s/	/sh/
/s/	95.66	4.34
/sh/	2.60	97.40

Table 9: Confusion matrix showing % classification accuracy of classification

Identified → / Actual ↓	/s/	/sh/
/s/	**96.64**	3.36
/sh/	1.34	**98.66**

Table 10: Confusion matrix showing % classification accuracy of unvoiced fricative classification using MFCC

Identified → / Actual ↓	/f/	/th/	/s/	/sh/
/f/	84.95	7.64	3.8	2.64
/th/	24.49	56.49	10.56	7.04
/s/	2.39	2.98	92.69	1.94
/sh/	2.28	1.41	1.69	94.60

Table 11: Confusion matrix showing % classification accuracy of unvoiced fricative classification using CFCC

Identified → / Actual ↓	/f/	/th/	/s/	/sh/
/f/	**82.83**	15.19	1.9	1.15
/th/	22.63	**72.50**	2.48	0.98
/s/	0.625	1.8	**95.77**	1.81
/sh/	1.62	0.82	1.73	**95.83**

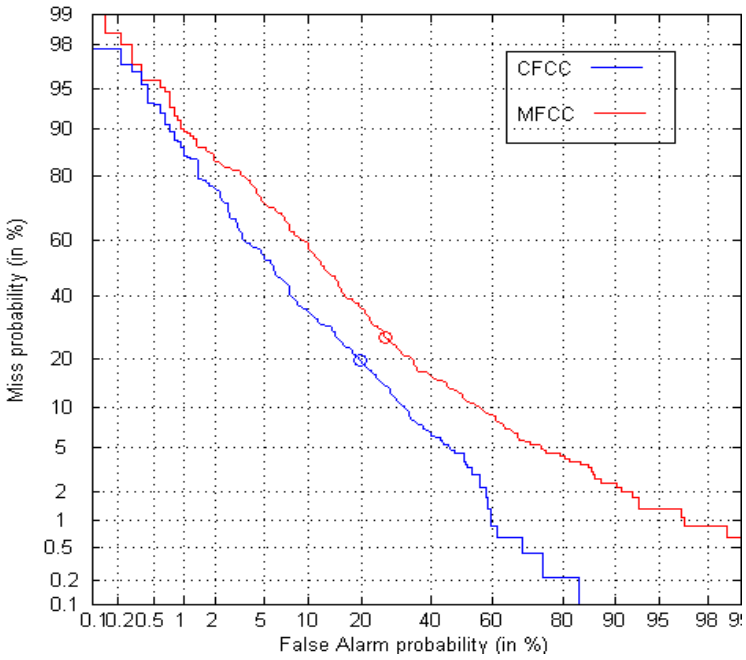


Fig.11 DET curves for unvoiced fricative classification (for *four* classes, *viz.*, /f/, /th/, /s/ and /sh/) using MFCC and CFCC.

5.2. Analysis of data independency via *4*-fold cross-validation

Classification results should *not* be data-dependent(*i.e.,* specific to particular set of training and testing samples)rather should be consistent for any dataset as long as datasets are valid (*i.e.,* represent samples from respective classes).In this paper, this is ensured by evaluating classification results using *4*-fold cross-validation analysis. Data for each fricative class is *randomly* divided into *4* sets (as shown in Table 12) and each dataset is used for testing at a time while remaining datasets are used for training. Four such trials have been performed and corresponding experimental results for individual fricative classification are shown in Table 13 and Fig. 12. Table 13 shows the overall classification results for each fold while results for each fricative (averaged over all these *4* folds datasets) have been shown in Fig. 12.

CFCC proves to be a better front-end feature set for classification as training and testing datasets are varied in each of *4* folds (as shown in Table 13). It is also clear that both % EER and minimum DCF have been reduced in *4*-fold cross-validation analysis with slight reduction in accuracy as well compared to *2*-fold cross-validation analysis performed in Section 5.1 (as shown in Table 3). One of the possible reasons for this difference in results could be the trade-off involved between number of training and testing samples. Only half of the total samples have been used for training in *2*-fold cross-validation analysis whereas *75* % of total samples are used for training in case of *4*-fold cross-validation leading to better estimation of HMM parameters.

Table 12. Division of database into four sets via *4*-fold cross-validation

	# of test samples in *Dataset 1*	**# of test samples in *Dataset 2***	**# of test samples in *Dataset 3***	**# of test samples in *Dataset 4***	**Total # of samples**
/f/	52	52	52	52	208
/th/	36	35	36	36	143
/s/	77	76	76	76	305
/sh/	63	64	63	64	254

Table 13. % classification accuracy for different training and testing sets (using *4*-fold cross-validation) using CFCC and MFCC for classification of /f/,/th/, /s/ and /sh/

Fold number	**Fold-*1***		**Fold-*2***		**Fold-*3***		**Fold-*4***		**Average**	
Feature set → / Results ↓	MFCC	CFCC	MFCC	CFCC	MFCC	CFCC	MFCC	CFCC	MFCC	CFCC
% classification accuracy	84.11	**87**	83.67	**89.13**	85.93	**90.28**	87.58	**85.65**	85.32	**88.01**
% EER	24.64	**18.07**	27.75	**18.94**	24.86	**16.75**	26.60	**19.60**	25.96	**18.34**
Minimum DCF	0.1937	**0.148**	0.209	**0.141**	0.2050	**0.141**	0.196	**0.151**	0.2010	**0.145**

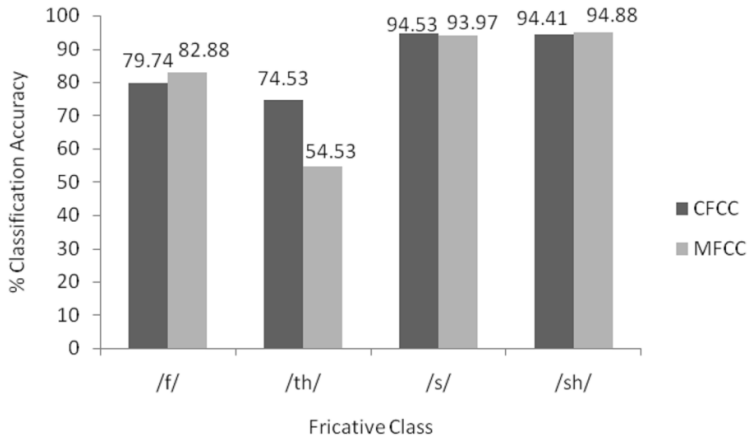
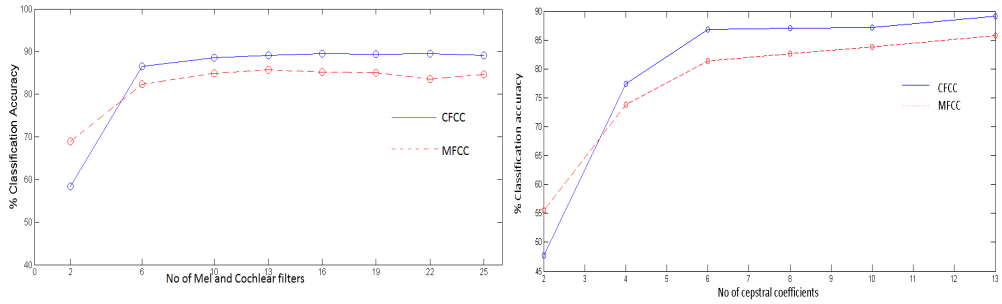


Fig. 12. *4*-fold averaged classification accuracy for individual fricative class for classification of /f/, /th/, /s/ and /sh/ using CFCC.

However, this is accompanied by less conclusive classification analysis as testing samples have been reduced. Averaged individual fricative class accuracy (as shown in Fig. 12) shows significant difference in accuracy of non-sibilant, /th/ using CFCC and MFCC features (*i.e.,74.53 %* for CFCC, *54.53 %* for MFCC) confirming *dataset independence* of experimental results reported in Section 5.1.

5.3. Effect of number of sub-band filters and feature dimensions

Both proposed (CFCC) and baseline (MFCC) features are evaluated by applying different number of sub-band filters on corresponding spectral information to estimate optimum number of Mel and cochlear filters required to capture distinct acoustic characteristics of each class. In wavelet analysis, there is always a trade-off between number of sub-band filters used and associated computational complexity. As more number of filters tend to provide more resolution (both in time and frequency-domain), it is intuitive that this number should be chosen based on a particular application (*i.e.,* minimum number providing sufficient *temporal* and *spectral* details). Initially, we varied the number of sub-band filters used to estimate feature vector along with dimensions of feature set. In particular, if number of sub-band filters used is *N* then dimension of feature vector is also kept as *N*. Fig. 13 (a) shows the plot of *%* classification accuracy *vs.* number of sub-band filters (with fixed feature dimension) whereas Fig. 13(b) shows the plot of *%* classification accuracy *vs.* dimension of feature vector (with fixed number of sub-band filters)
.



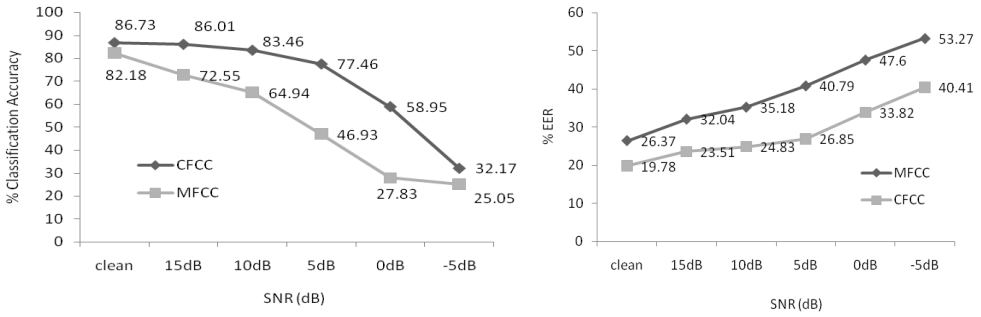
(a) (b)

Fig. 13 (a) % classification accuracy with variation in number of filters employed and with fixed dimension of feature vector for classification of /f/, /th, /s/ and /sh/, (b) % classification accuracy by varying feature dimension and with fixed number of filters for classification of /f/, /th/, /s/ and /sh/.

Feature dimension of *13* (with *13* cochlear sub-band filters) is found to be *optimum* for both CFCC and MFCC features as both features show near -maximum classification accuracy in (*i.e., 89.14 %* for CFCC, *85.73 %* for MFCC when number of filters are varied by keeping fixed dimension of feature vector). Hence, all the other experiments reported in this work have been performed using *13*-dimensional feature vectors for both CFCC and MFCC. In the next experiment, we fixed the number of sub-band filters and reduced the number of cepstral coefficients (*i.e.,* feature dimension) from *13* in order to examine how many cepstral coefficients are vital. Fig. 13 (b) shows the results obtained as feature dimensions are varied alone (*i.e.,* with fixed number of sub-band filters). It is observed that employing only *6* cepstral coefficients of CFCC results in considerable classification accuracy in both the experiments (*i.e., 86.48 %* when *6* filters are employed, and *86.77 %* when number of filters are fixed to *13*) followed by rapid fall in accuracy on reducing the feature dimension further. Therefore, it can be concluded that these *6* cochlear filters provide enough spectral resolution for capturing the distinctive spectral characteristics of given unvoiced fricatives. Impulse and frequency responses of these *6* cochlear filters have been discussed in Section 3 (as shown in Fig. 3 and Fig. 4, respectively).

5.4. Robustness under signal degradation conditions

To study the robustness of the proposed feature set under noisy conditions, testing samples of fricative sounds were *added* with white noise at various SNR levels, while training is performed with clean fricative samples. White noise samples are obtained from *NOISEX-92* database [29] (having sampling frequency of *19.98* kHz). These noise samples have been up-sampled to *44.1* kHz such that up-sampled white noise contains all the frequencies up to *22.05* kHz. Analysis is performed on these test samples using both MFCC and CFCC features starting from clean conditions and at varying SNR levels from *15* dB to *-5* dB in steps of *5* dB. Fig. 14 shows the performance of both features under various SNR levels. Though overall classification accuracy decreases in case of both features, the decrease is much *steeper* with MFCC features, as accuracy falls to *46.93 %* at SNR of *5* dB while CFCC accuracy still remains at *77.46 %*. Similar behavior has been observed in % EER as well since % EER has been considerably increased with SNR degradation in case of MFCC (*i.e., 26.37 %* EER in clean conditions to *40.49 %* EER at SNR of *5* dB) while this increase is less steeper in CFCC (*i.e., 19.78 %* EER in clean conditions to *26.85 %* EER at SNR of *5* dB).



(a) (b)

Fig. 14. (a) Degradation of average classification accuracies in presence of additive white noise using baseline (MFCC) and proposed (CFCC) feature sets, (b) increase in classification error in presence of white noise using baseline (MFCC) and proposed (CFCC) feature sets.

As discussed in Section 3.3, the robustness of CFCC is due to the fact that

1. CFCC employs smooth bell-shaped cochear filters as opposed to triangular-shaped Mel filters,

2. CFCC is designed to mimic human auditory processing which has inherent noise suppression mechanism to take care of mechanical vibration of eardrum at the threshold of hearing,
3. CFCC employs CWT which has mother wavelet to aid the noise suppression in wavelet-domain.

Decreasing SNR levels beyond *5* dB SNR results in rapid fall of accuracies in case of both feature domains as fricative sounds are almost *masked* by added white noise and front end features no longer reflect distinct acoustic characteristics in presence of such high noise.

6 SUMMARY AND CONCLUSIONS

Application of recently developed auditory-based cochlear filters for identifying spectral cues in unvoiced class of fricatives has been proposed. Study was motivated by need to develop effective acoustic cues using auditory transform pertaining to the similarity of auditory transform with human cochlear response thereby distinguishing effectively between fricative sounds. Our experimental results indicate that proposed CFCC features outperform MFCC features both in clean and noisy conditions. One of the possible limitations of this study could be classification is solely dependent on spectral characteristics of *manually* segmented fricative sounds. Including contextual information may result in better classification since proposed feature set, *viz.,* CFCC itself depends on human auditory system and contextual information greatly helps in perceiving fricative utterances in case of human listeners[30]. Global optimization of HMM parameters is another issue as Baum-Welch re-estimation algorithm guarantees only local optimization.

Auditory transform-based CFCC features present an alternative to state-of-the-art front end features (*viz.,* MFCC) used for robust phoneme classification. Our future research will be directed towards extending our present study to application of proposed robust feature (*i.e.*, CFCC) in phoneme identification task.

REFERENCES:

[1] Fant, G., Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[2] Stevens, K.N., Acoustic Phonetics (Current Studies in Linguistics), M.I.T. Press, 1999.

[3] J. D. Markel and A. H. Gray Jr., Linear Prediction of Speech, Springer-Verlag, 1976.

[4] Q. Li, An auditory-based transform for audio signal processing, Proc. IEEE Workshop App. Signal Process. Audio Acoust., New Paltz, NY, pp. 181–184, Oct. 2009.

[5] Qi Li, An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions, IEEE Trans. on Audio, Speech and Lang. Process., vol. 19, no. 6, pp.1791-1801, Aug. 2011.

[6] McCasland, G. P., Noise intensity and spectrtuirt cues for spoken fricatives, J. Acoust. Soc. Am. Suppl. vol. 165, pp.S78–S79, 1979.

[7] Behrens, S. and S. E. Blumstein, Acoustic characteristics of English voiceless fricatives: a descriptive analysis, J. Phonetics, vol. 16, no.3, pp. 295–298, 1988.

[8] Stevens, K. N., Evidence for the role of acoustic boundaries in the perception of speech sounds, J. Acoust. Soc. Am., vol. 69, no. S1, pp. S116-S116, 1981.

[9] Behrens, S. and S. E. Blumstein, On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants, J. Acoust. Soc. Am., vol. 84, no. 3, pp. 861–867, 1988.

[10] Jongman, A., Duration of fricative noise required for identification of English fricatives, J. Acoust. Soc. Am., vol. 85, no. 4, pp. 1718–1725, 1989.

[11] Jongman, A., R. Wayland, and S. Wong, Acoustic characteristics of English fricatives, J. Acoust. Soc. Am., vol. 108, no.3, pp. 1252–1263, 2000.

[12] Hughes, G. W. and M. Halle, Spectral properties of fricative consonants, J. Acoust. Soc. Am., vol. 28, no.2, pp. 303–310, 1956.

[13] Strevens, P., Spectra of fricative noise in human speech, Lang. Speech, vol. 3, no.1, pp. 32–49, 1960.

[14] Pentz, A., H. R. Gilbert, and P. Zawadzki, Spectral properties of fricative consonants in children, J. Acoust. Soc. Am., vol. 66, no. 6, pp. 1891–1893, 1979.

[15] Nissen, S., An accoustic analysis of voiceless obstruents produced by adults and typically developing children, Ph. D. Thesis, Ohio State University, Columbus, OH, 2003.

[16] S. Mallat, A Wavelet Tour of Signal Processing, 3rd Ed., New York: Academic, 2007.

[17] Qi Li, An auditory-based transform for audio signal processing, IEEE workshop on applications of signal processing to audio and acoustics – WASPAA, pp. 181-184, 2009.

[18] J.L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Ed., Springer-Verlag, New York, 1972.

[19] R.K. Potter, G.A. Kopp, and H.C. Green, Visible Speech, D.Van Nostrand Co., New York, 1947. Republished by Dover Publications, Inc., 1966.

[20] S. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. on Acoustics, Speech and Signal Process., vol. 28, no 4, pp. 357-366, Aug. 1980.

[21] H. M. Teager and S. M. Teager, Evidence for nonlinear production mechanisms in the vocal tract, in Speech Production and Speech Modeling, Norwell, MA: Kluwer, vol. 55, pp. 241–261, 1989.

[22] Brownell, W. E., Bader, C. R., Bertrand, D. and Ribaupierre, Y. d. , Evoked Mechanical Responses of Isolated Cochlear Outer Hair Cells, Science, vol. 227, pp. 194-196, 1985.

[23] Helmholtz, H. L. F. v. On the Sensations of Tone, Dover Publications, Inc., New York, NY, 1954.

[24] S. Mallat, A Wavelet Tour of Signal Processing, 3rd Ed., New York: Academic, 2007.

[25] Jaffard, Pointwise smoothness, two-microlocalization and wavelet coefficients. Publications Mathematiques, vol. 35, pp. 155168, 1991.

[26] Audacity software: Available Online: http://audacity.sourceforge.net/ {Last accessed : July 22, 2013}.

[27] CHAINS Corpus: Available online: http://chains.ucd.ie/ftpaccess.php .{Last accessed : July 22,2013}.

[28] Marija Tabain and Catherine Watson, A study on classification of fricatives,6th Australian International conference on Speech science and technology, Adelaide, pp. 623-628, Dec.1996

[29] White Noise Source: NOISEX-92 database , Available online : http://spib.rice.edu/spib/data/signals/noise/white.html {Last Accessed : July 22, 2013}.

[30] Brian C. J. Moore, An Introduction to the Psychology of Hearing, Academic Press, 4th Ed., 1997.

[31] Frid,A., Lavner,Y., Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm, 26th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI'10), pp.751-755, 2010.

[32] A.F. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, The DET curve in assessment of detection error performance, Proc. EUROSPEECH'97, Rhodes Greece, vol.4, pp.1899-1903, Sept. 1997.

[33] Bolle R.M., Pankanti S., Ratha N.K., Evaluation techniques for biometrics-based authentication systems (FRR), Proc. 15th International Conference on Pattern Recognition , vol.2, pp.831-837, 2000.