RI 11009, 6 June 2011 Computer Science

# IBM Research Report

## Fepstrum Features: Design and Application to Conversational Speech Recognition

IBM Research Report No. RI 11009
6th June 2011

**Vivek Tyagi**
IBM Research Division
IBM India Research Lab
Plot 4, Block-C, Vasant Kunj
New Delhi - 110070, India.

# Fepstrum Features: Design and Application to Conversational Speech Recognition

Vivek Tyagi, *Member, IEEE*

***Abstract*—In this paper, we present the Fepstrum features – a principled approach to estimate the modulation spectrum of the speech signals using the Hilbert envelopes in a non-parametric way. The importance of the modulation spectrum as a feature in the automatic speech recognition (ASR) has long been established by several researchers in the past two-three decades. However, traditionally, in the speech recognition literature the modulation spectrum features have been extracted as the DCT/DFT of the $\log$ Mel filter's energies over $10-15$ frames. These Mel-filter energies are in-turn computed through short term spectrum (with $20-30ms$ long primary window). We show, that this approach leads to a crude approximation of the modulation spectrum in the Mel-filter bands. Further, we show that the $\log$ of a particular Mel-Filter's Hilbert envelope (obtained over a primary analysis window of $100ms$) leads to a principled amplitude modulation (AM) signal estimate in that band. Lower DCT coefficients (in the range $0-25Hz$) of the AM signal leads to the *fepstrum* features. To assess the effectiveness of the *fepstrum* features, we have performed conversational telephony speech (CTS) recognition experiments on the Switchboard (SWB) corpus using a recently developed LVCSR library (IBM IrlTK). Our experiments indicate that the *fepstrum* features in simple concatenation with the short-term spectral envelope features (MFCC) provide upto $2.5\%$ absolute improvement in phoneme recognition accuracy and upto $2.5\%-3.5\%$ absolute word recognition accuracy improvement on a $1.5Hr$ SWB test set with a $2,300$ words vocabulary. We also provide the details of our IrlTK LVCSR acoustic modeling library.**

**EDICS: SPE-ANLS,SPE-RECO,SPE-LVCR**

## I. Introduction

Amplitude Modulation (AM) is a very efficient and simple technique in the electronic communication systems for the transmission of the "information bearing" base-band signals over a carrier signal (usually a monochromatic sinusoid)[46]. It has been in use for over a century now owing to its simple mathematical characterization when the carrier signal is a single sinusoid as it leads to a simple demodulation scheme at the receiver end. While the original AM formulation with sinusoidal carriers was done in the field of electronic communication systems[46], it turns out that the natural base-band signals such as speech and music can also be faithfully represented by a sum of low frequency modulators which modulate higher frequency carriers[1], [4], [5], [6], [7], [8], [11]. In this case, however, the carriers are not monochromatic sinusoids but are non-zero finite band-width carriers whose central frequencies also change over time.

Several studies have shown that the amplitude modulators (AM) of a speech signal are very important for the speech perception[4], [5], [6], [7]. For example, when the AM signals of a speech signal are replaced by the envelopes of constant magnitude, speech becomes unintelligible. However, speech remains highly intelligible when the AM signals are preserved, but the carriers are replaced by the white noise. Therefore, amplitude modulator or equivalently its modulation spectrum (spectrum of the AM signal) has been an area of great interest for the speech recognition and speech processing/filtering. Any modulation analysis technique entails separation of a broad-band signal (such as speech, music) into several narrow-band frequency sub-bands followed by the amplitude modulation and frequency modulation (AM-FM) decomposition in the corresponding sub-band. Several such methods are described in literature[13], [14], [15], [16], [18], [23], [25], but the details of how a signal is separated into sub-bands, and how a sub-band is decomposed into a carrier and modulator vary from the author to author. Most of these techniques define and extract the modulation spectrum in a slightly ad-hoc manner. For instance, several researchers have extracted the speech modulation spectrum by computing a discrete Fourier transform (DFT)/DCT of the Mel or critical band spectral energy trajectories, where each sample of the trajectory has been obtained through a short-term spectrum (followed by the Mel filtering) over $20-30ms$ long windows. An illustration of this is provided in Fig.1. There are two major limitations of such techniques. First, they implicitly assumes that within each Mel or critical band, the amplitude modulation (AM) signal remains constant throughout the duration of the primary window length that is typically $20-30ms$ long. Second, instead of modeling the constantly and slowly changing amplitude modulation signal in each band, they mostly model the spurious and abrupt modulation frequency changes that occur due to the frame shifting of $10ms$.

Furthermore, these techniques only provide feature extraction for the speech recognition and therefore do not concern themselves with the carrier signal extraction and/or signal filtering and signal reconstruction which would have required a more principled approach.

In the recent past, several researchers, notably Atlas, Schimmel, Sukitannon [8], [9], [10], [11] and Kumaresan and Rao[1], [2], [3] have been developing a unified and formal study of the modulation analysis and filtering systems from a signal processing perspective and not just the modulation feature extraction for the speech recognition. In particular, Atlas, Janssen and Schimmel have argued for the need for a "coherent" AM-FM decomposition[8], [9] in order to achieve meaningful and artifact-free modulation filtering. Margos et. al[13] have proposed the use of the Teager-energy operator for

V. Tyagi is with IBM India Research Lab, New Delhi (email: vivetyag@in.ibm.com)

the AM-FM demodulation of the speech signals.

In a series of papers, Kumaresan and Rao have developed an auto-regressive parametric technique to decompose a speech signal into several narrow-bands followed by a principled AM-FM decomposition in each sub-band[1], [2], [3]. They show that the analytic signal of a speech signal can be modeled as a conventional pole-zero model in the time domain, just as the spectral envelope of speech and several other signals are modeled by an all-pole model in the frequency domain[44], [45]. Through a series of simplifications, they show that the AM signal can be estimated by solving for the linear prediction (LP) equations, though in the frequency domain, much in the same way as the classical LP spectral envelope[44], [45] of a speech signal is estimated by solving for the linear prediction equations in the time domain. Therefore Kumaresan et. al. have also termed their technique as linear prediction in spectral domain (LPSD). Following up with their work Athineos et al. [23], [24] have used the dual of the linear prediction in the spectral domain to improve upon the TRAP[16] features. However, as is well known, any all-pole model suffers from two major limitations[12],

- A suitable model order of the all-pole filter has to be empirically chosen to estimate the envelope of the underlying signal. Further, the model order may change from frame to frame as the underlying speech signal's characteristics change. However, in practice, a fixed model order is chosen while we clearly realize that it may be sub-optimal.
- At the location of the poles, the all-pole envelope (whether in the time or frequency domain) becomes undefined as it takes the undefined form of $\frac{1}{0}$. Further, due to the fixed numerical accuracy of the digital computers, this fraction (all-pole envelope) evaluates to a random high number in the vicinity of the pole. This fact has been well described by Murthi and Rao[12] where they propose the use of a minimum variance distortion-less spectrum (MVDR) instead of the all-pole model based LP spectrum to alleviate this problem.

Therefore, in this paper, we extend the approach of Kumaresan et al. by estimating the AM signal in each sub-band. However, unlike their work, we will estimate the AM signals in a *non-parametric* way, without using the LP equations in the spectral domain which may suffer from the above listed limitations. We propose an algorithm to perform the AM-FM demodulation of the speech signal in the time domain. As the AM-FM signal model is defined in time domain[1], a demodulation in time domain leads to conceptual clarity and a better understanding of the relationships between various signal sub-components.

In this paper, we denote the well studied pole-zero (LP) models that are used for modeling the short-term spectrum[44], [45] of a signal, by *"F-PZ"*. Lately, Kumaresan and Rao[1], [2], [3] have proposed to model analytic signals[1] using the pole-zero models in the temporal domain (denoted by *T-PZ* to distinguish them from the *F-PZ*).
An inherent advantage of working with the analytic signal is that it elegantly allows the decomposition of an arbitrary signal (possibly non-stationary) into its amplitude modulation (AM) and frequency modulation (FM) signals. We make extensive use of *T-PZ* representation in this paper. For the sake of completeness and clarity, we will describe and prove several interesting time-frequency dualities for the analytic signals. These properties are then used to develop "meaningful" AM-FM decomposition of the speech signal in a non-parametric way.

We have termed this approach to extract the AM signal and its modulation spectrum, as the *fepstrum*. We later show in the paper that the *fepstrum* is a signal processing dual of the well-known quantity *real cepstrum*. Furthermore, we evaluate the efficacy of the fepstrum features on the large vocabulary speech recognition experiments based on the Switchboard (SWB) conversational telephony speech (CTS) corpus. At the IBM India Research Lab (IRL), we have developed a C++ based LVCSR acoustic modeling library (IrlTK) and a beam pruned, time synchronous, Viterbi LVCSR recognizer that uses the IrlTK trained acoustic models on the $70Hr$ of the Switchboard data. Our LVCSR recognition results indicate that the fepstrum feature, in simple concatenation with the short term spectral envelope features (MFCC), provides upto $2.5\% - 3.5\%$ absolute word recognition accuracy improvement on a $1.5Hr$ SWB test set. We attribute this improvement to the complementary information present in the fepstrum features. While MFCC provides short term spectral envelope feature that is well localized over $20 - 30ms$ of the speech signal, fepstrum feature provides principled modulation spectrum of the slowly moving AM signal in each Mel-band, over $100ms$ of the speech signal.

This remainder of this paper is organized as follows. In section II, we describe the use of pole-zero models to represent the speech analytic signal in the time domain and show its factorization in terms of the AM and FM signals of each sub-band. In section III, we describe a non-parametric technique to estimate the modulation spectrum of the AM signals in each sub-band and term it as *fepstrum*. In section IV we describe our Switchboard (SWB) experimental setup, a new LVCSR acoustic modeling library (IrlTK) followed by the detailed experimental results. Finally, we summarize our findings in section V.

## II. Pole-zero models in the temporal domain

Traditionally, the pole-zero transfer functions have been used to approximate a discrete time frequency response which is inherently periodic with a period of $2\pi$. Voelcker and Kumaresan have used the T-PZ to approximate analytic signals in the temporal domain. We recall that given a real periodic[2] signal $x(t)$ with period T seconds, its analytic version $s(t)$ is given by,

$$s(t) = x(t) + j\hat{x}(t) \tag{1}$$

[1] $x(t) = a(t)cos(\int_0^t 2\pi f(t)dt)$, here $x(t)$ is a narrow band-pass filtered speech signal where, $a(t)$ is the corresponding AM signal and $f(t)$ is the corresponding FM signal

[2] This is not a limitation as in short-time Fourier analysis, we implicitly make the signal periodic with the base period equal to the $T$ second long windowed segment.

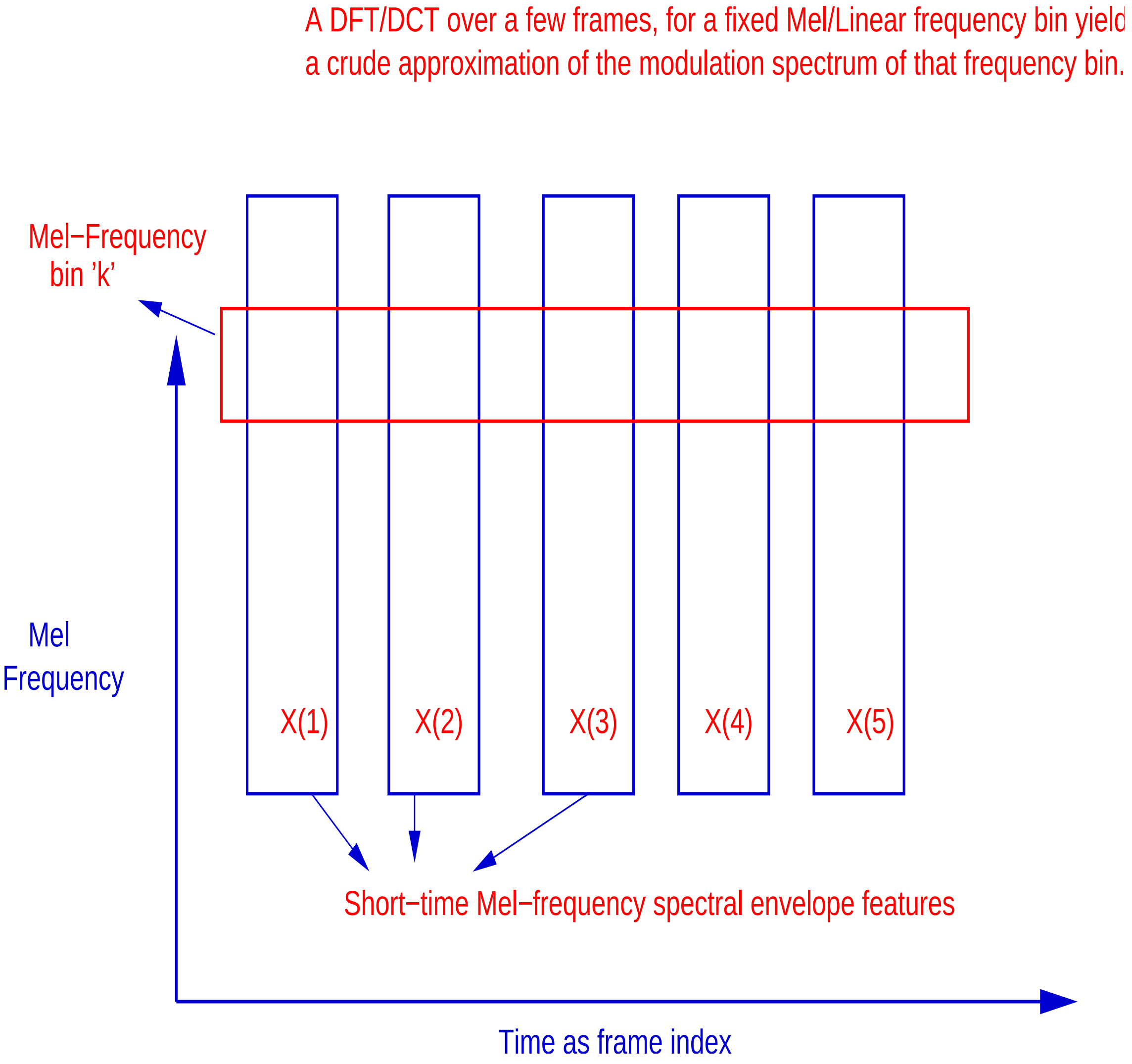
A DFT/DCT over a few frames, for a fixed Mel/Linear frequency bin yielc
a crude approximation of the modulation spectrum of that frequency bin.
Mel−Frequency
bin 'k'
Mel
Frequency
X(1)
X(2)
X(3)
X(4)
X(5)
Short−time Mel−frequency spectral envelope features
Time as frame index

Fig. 1. An approximate AM demodulation technique.

where $\hat{x}(t)$ denotes the Hilbert transform of $x(t)$. If $x(t)$ is band-limited, then so is $s(t)$. Moreover $s(t)$ has non-zero spectrum for only positive frequencies. Therefore $s(t)$ can be expressed in terms of a finite number of Fourier series coefficients at positive frequencies.

$$s(t) = e^{j\omega_t t} \sum_{k=0}^{M} a_k e^{jk\Omega t} \qquad (2)$$

where $\omega_t$ is an arbitrary frequency translation, $\Omega = 2\pi/T$ and $M$ is sufficiently large. Noting that $s(t)$ is a polynomial, it can be factored in terms of T-PZ as follows,

$$s(t) = a_0 e^{j\omega_t t} \prod_{i=1}^{P}(1 - p_i e^{j\Omega t}) \prod_{l=1}^{Q}(1 - q_l e^{j\Omega t}) \qquad (3)$$

where $P+Q = M$ and $p_i$ and $q_l$ are the complex roots, inside and outside the unit circle respectively. We note that this is a unit circle in the time domain, $e^{j\Omega t}$, $t \in [0, T]$, $\Omega = \frac{2\pi}{T}$. More generally, if s(t) is not band-limited, it can be represented using poles and zeros.

$$s(t) = a_0 e^{j\omega_t t} \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \prod_{l=1}^{Q}(1 - q_l e^{j\Omega t}) \qquad (4)$$

where, $p_i$ and $q_i$ are the zeros inside and outside the unit circle respectively. The poles $u_i$ are guaranteed to be inside the unit circle as proved in the following lemma.

**Lemma 1** *The T-PZ factorization of an analytic signal $s(t)$ has all the poles $u_i$ inside the unit circle.*

Proof: Lets assume that there is a pole $r = |r|e^{j\phi}$ outside the unit circle, with $|r| > 1$. The expansion of $s(t)$ will then have a term,

$$\begin{aligned} \frac{A}{(1 - re^{j\Omega t})} &= \frac{-A}{re^{j\Omega t}} \frac{1}{1 - r^{-1}e^{-j\Omega t}} \\ &= \frac{-A}{re^{j\Omega t}} \sum_{k=0}^{\infty} r^{-k} e^{-jk\Omega t} \end{aligned} \qquad (5)$$

where, A is a constant. (5) implies that $s(t)$ has non-zero spectrum for negative frequencies. This is in contradiction to the fact that $s(t)$ being an analytic signal has zero spectral energy for negative frequencies. Hence $|r| < 1$.

The importance of lemma 1 will become apparent later on. Let us now specify the dual analogues of three well known properties which are,

- Minimum-phase: Traditionally, minimum phase is a frequency domain phenomenon. A frequency response (F-PZ) is termed minimum-phase (F-MinP) if all its poles and zeros are inside the unit circle. Similarly, a T-PZ is called T-MinP if all its poles and zeros are inside the unit circle.
- All-pass: Traditionally, all-pass is a frequency domain phenomenon. A frequency response, (F-PZ), is said to be all-pass (F-AllP) if its magnitude is unity at all frequencies. Similarly, a T-PZ is called T-AllP if it has unity magnitude for $t \in (-\infty, \infty)$.
- Causality: Traditionally, causality is a time-domain phenomenon. A signal $x(t)$ is said to be causal (T-causal) if it is non-zero only for the $t \geq 0$. Similarly, we define a frequency response to be F-causal if it is non-zero only for the $f \geq 0$. Therefore, an analytic signal is F-causal.

With these definitions in place, we are ready to describe the decomposition of an analytic signal $s(t)$ into its T-MinP and T-AllP part which will lead to its AM and FM parts. Therefore, reflecting the zeros $q_i$ inside the unit circle, we get,

$$s(t) = a_0 e^{j\omega_t t} \underbrace{\frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \prod_{i=1}^{Q}(1 - 1/q_i^* e^{j\Omega t})}_{\textbf{T-MinP}} \qquad (6)$$

$$\times \prod_{i=1}^{Q}(-q_i^*) \underbrace{\prod_{i=1}^{Q} \frac{(e^{-j\Omega t} - q_i)}{(1 - q_i^* e^{-j\Omega t})}}_{\textbf{T-AllP}}$$

We recall the following two well-known lemmas,

**Lemma 2** *Given a frequency response (F-PZ) $X(f)$ $= |X(f)|e^{j\phi(f)}$, its phase response $\phi(f)$ is the Hilbert transform of its log-envelope $\log|X(f)|$, if and only if the frequency response is minimum phase (i.e a F-PZ with all the poles and zeros inside the unit circle).*

**Lemma 3** *Given a frequency response (F-PZ) $X(f)$ $= |X(f)|e^{j\phi(f)}$, it is minimum phase, if and only if, its complex cepstrum (CC) $x_{cc}(n)$ is causal (i.e $x_{cc}(n) = 0, n \in [-\infty, -1]$)*

The proof of above two lemmas can be found in the pages 782-783 of [45]. Using the *time-frequency* duality, we will state and prove a dual of the lemmas (2), (3).

**Lemma 4** *Given an analytic T-PZ signal $s(t)$ $= \frac{\prod_{i=1}^{P}(1-p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1-u_i e^{j\Omega t})} = |s(t)|e^{j\Psi(t)}$, all of its poles and zeros are within the unit-circle (i.e s(t) is T-MinP) if and only if its phase $\Psi(t)$ is the Hilbert transform of its log envelope $\log|s(t)|$.*

Proof: Let $\tilde{S}(f)$ be the Fourier transform (FT) of $\log s(t) = \log|s(t)| + j\Psi(t)$. We note that $\tilde{S}(f)$ consists of spectral lines at integral multiple of $\Omega$[3] and hence is a discrete sequence. Let us assume that the phase $\Psi(t)$ is the Hilbert transform of the log envelope $\log|s(t)|$. This implies that $\log s(t)$ is an analytic signal and hence its FT $\tilde{S}(f)$ is zero for negative frequencies (i.e. $\tilde{S}(f)$ is a discrete and F-causal sequence). Using the duality principle we note that $\log s(-f)$ is the FT of $\tilde{S}(t)$. In fact, $\tilde{S}(t)$ is the complex cepstrum(CC) of a signal whose FT is $s(-f)$. As $\tilde{S}(t)$ has the same functional form as $\tilde{S}(f)$, this implies that $\tilde{S}(t)$ is a discrete and causal CC sequence. Therefore in light of lemma (3), it follows that $s(-f)$ is minimum-phase F-PZ with all the zeros and poles inside the unit circle. Therefore we get,

$$s(-f) = \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega(-f)})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega(-f)})}$$

substituting t for '-f' we get,

$$s(t) = \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \qquad (7)$$

This proves that the T-PZ $s(t)$ that is T-MinP results in its phase being the HT of its log-envelope.

Therefore, using Lemma (4), $s(t)$ can be expressed as follows,

$$s(t) = \underbrace{a_0 \prod_{i=1}^{Q}(-q_i^*)}_{A_c} \underbrace{e^{\alpha(t) + j\hat{\alpha}(t)}}_{\textbf{T-MinP}} \underbrace{e^{j\gamma(t)}}_{\textbf{T-AllP}} \qquad (8)$$

where $A_c$ is a constant, $\alpha(t)$ is the logarithm of the absolute magnitude of the AM signal, $\hat{\alpha}(t)$ its HT and $\hat{\alpha}(t) + \gamma(t)$ is the phase signal and its derivative is the FM signal.

$$\log|s(t)| = \alpha(t) + \log(A_c) \qquad (9)$$

$$(10)$$

As, $\hat{\alpha}(t)$ can be determined from the log AM signal $\alpha(t)$ [4], it forms the redundant information and hence is excluded from the FM signal. Therefore, $\gamma'(t)$ is the FM (instantaneous frequency) signal of interest, where $'$ denotes derivative with respect to time.

The next step is to develop an algorithm that can automatically achieve the decomposition as in (8). Noting that the all-pole F-PZ as estimated using classical linear prediction technique is guaranteed to be minimum phase, Kumaresan et. al. used the dual of linear prediction in the spectral domain (LPSD) [1], [2], with sufficiently high prediction order 'M', to derive the T-MinP signal. The T-AllP signal was obtained as the residual signal of the LPSD.

However, it is well known that the LP technique overestimates the peaks and poorly models the valley. In particular, for values close to the poles, the all-pole model takes up an indefinite form $\frac{1}{0}$[12]. However, due to the fixed numerical accuracy of the digital computers, this fraction (all pole envelope) evaluates to a random high number in the vicinity of the pole leading to erroneous envelope estimate in that

[3] This can be seen by series expansion of $\log(1 - pe^{j\Omega t}) = \sum_{k=1}^{\infty} -p^k e^{jk\Omega t}/k$

[4] Due to the HT relationship between the two

vicinity[12]. Moreover, the results are highly susceptible to the model order 'M' whose actual value is not known.

Therefore, in this work, we use a non-parametric technique to estimate the AM signals. From (8), we note that $\log|s(t)| = \alpha(t) + \log(A_c)$, where $\log(A_c)$ is a constant over the frame. Therefore the logarithm of the absolute magnitude of the analytic signal in each band is an estimate of the corresponding AM signal + a constant term. In the following section, we explain how to estimate the AM signal and its spectrum (modulation spectrum) in each sub-band in a non-parametric way. The corresponding FM signal $\gamma'(t)$ is extracted through homomorphic filtering and its details can be found in [20]. In the remainder of this paper, we are only interested in the AM signal for its use in the speech recognition as a feature vector.

## III. Fepstrum feature extraction

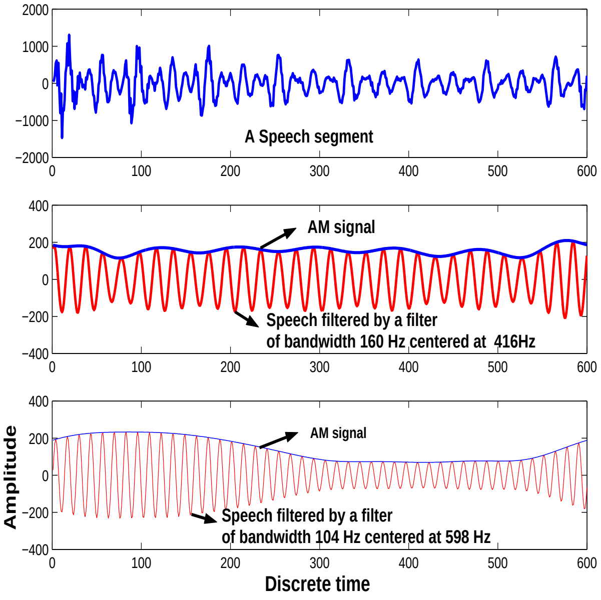

Fig. 3. The AM signal derived using narrow-band filters

Fig.2 illustrates our feature extraction scheme. A wide-band analytic speech signal s(t) (with typical bandwidth of $4KHz$ for telephony speech) is decomposed into $J = 24$ Mel-frequency sub-bands[5]. We take the log magnitude of the Mel-filtered analytic signal to obtain its corresponding AM signal $\alpha(t)$ as following (8) we have $\log|s(t)| = \alpha(t) + \log(A_c)$. At this point AM signal $\alpha(t)$ has the same sampling frequency as the original speech signal $s(t)$. However, the AM signals are slowly changing low frequency signals and hence they are appropriately low-pass filtered and then down-sampled by a factor of $40$. Following this, their lower DCT coefficients are retained as the feature vector. To distinguish this representation from the previous use of the word "modulation spectrum" [14], [15], [16], [18], [23], we have termed this representation as *fepstrum*. As is well known, real cepstrum is the inverse DFT of the $\log$ absolute magnitude spectrum of a signal. Whereas, fepstrum is the DFT/DCT of the $\log$ absolute magnitude of the analytic signal. Therefore, fepstrum is a dual of the well-known quantity: real cepstrum.

As is widely known, the adjacent frequency bands of the speech signal are correlated with each other. And thereby, the fepstrum of each Mel-band is also correlated with each other. Therefore, the fepstrum features from each of the Mel bands are concatenated together, and then de-correlated using a Principal Component Analysis transform (PCA or KLT). The PCA transform matrix is obtained by the SVD of the autocorrelation matrix of the concatenated fepstrum features from all the Mel-bands. We further reduce the dimensionality of the PCA transformed fepstrum feature by keeping only the coefficients corresponding to the top-few($50\%$) eigenvalues. These top $50\%$ eigenvalues accounted for almost $95.0\%$ of the total eigenvalues' mass.

Fig. 3 shows a particular example of our non-parametric AM signal estimation technique. First pane shows a wide-band speech signal which is then filtered by Mel-filters of bandwidth $160Hz$ and $104Hz$ respectively. Second and third pane shows the narrow band-pass filtered speech signals and their corresponding AM signals. We note that these AM signals are indeed low modulation frequency signals as they change rather slowly over time. Therefore, their lower DCT coefficients (fepstrum) lead to their parsimonious description. We next describe our SWB experimental setup.

## IV. Experiments and Results

### A. Switchboard train-test-dev sets

In order to assess the effectiveness of the fepstrum features, speech recognition experiments were conducted on the Switchboard-1 (SWB) corpus[35]. SWB corpus in one of the most important benchmarks for recognition tasks involving the large vocabulary conversational speech (LVCSR). It is a very challenging task notwithstanding the limitations posed by the telephone channel, including bandwidth, transducer, noise and echo. Fast speaking rates; poor co-articulation at the word boundaries; a wide range of dialects, speaking styles and accents; and the large variation in the pronunciations of the words, all present unique problems for the recognition of such spontaneous speech[36], [37]. Moreover, these conversations are heavily populated with dysfluencies such as ungrammatical pauses, stutters, laughter, repeats and self-repairs. The vocabulary is large and dominated by monosyllabic words which are typically hard to recognize[36].

For our experiments we took about first $72$ hours of speech data from SWB-1 (the data present in first $10$ CD-ROMs of the SWB-1 corpus) and divided it into disjoint test, development and train sets as follows

- **Test set:** Our test set consists of the first $13$ two-sided conversations in the SWB-1 corpus with the conversation id running from $SW2001$ through $SW2019$. It consists of about $1.5$ hours of speech, $900$ utterances and has about $15,500$ word-tokens in it. The vocabulary size is $2,300$ words.
- **Dev set:** consists of $100$ utterances that are disjoint from the test-set and was used to tune the word insertion

[5] The spectrum of the $100ms$ long analytic signal is multiplied with the triangular frequency responses of the Mel-filters, followed by the IDFT to obtain the Mel-filtered analytic signal of the corresponding Mel-band

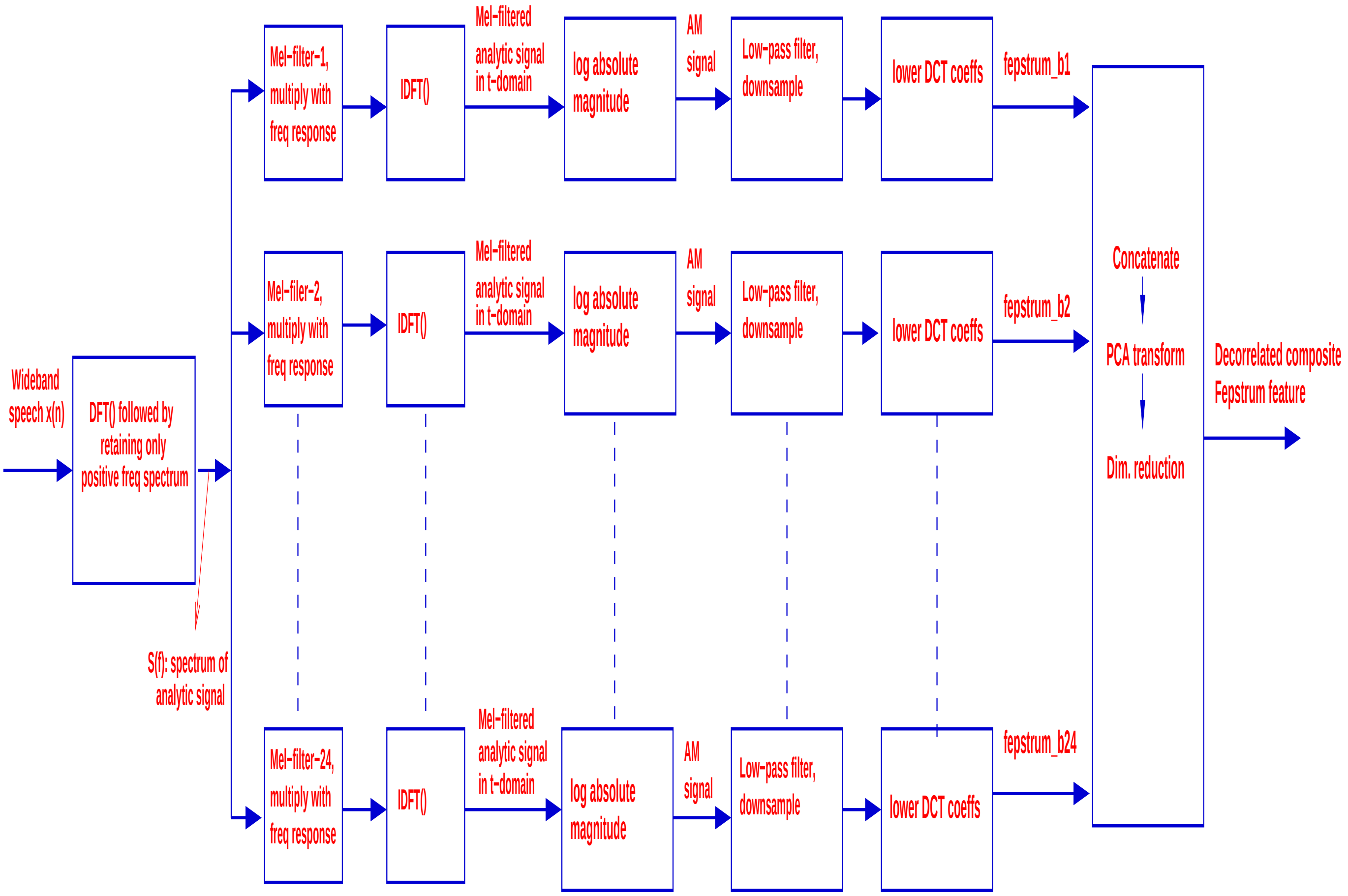
Wideband speech x(n)
DFT() followed by retaining only positive freq spectrum
S(f): spectrum of analytic signal
Mel-filter-1, multiply with freq response
Mel-filer-2, multiply with freq response
Mel-filter-24, multiply with freq response
IDFT()
Mel-filtered analytic signal in t-domain
log absolute magnitude
AM signal
Low-pass filter, downsample
lower DCT coeffs
fepstrum_b1
fepstrum_b2
fepstrum_b24
Concatenate
PCA transform
Dim. reduction
Decorrelated composite Fepstrum feature

Fig. 2. The FEPSTRUM feature extraction

penalty and the bigram language model weight factor used in the decoder for the recognition experiments.

- **Train set:** consists of the next 70 hours of the SWB-1 corpus and consists of about 700 two-sided conversations with the conversation id running from $SW2022$ through $SW2894$. The train set consist of about 1 million words.

### *B. LVCSR Library: IBM IrlTK*

At the IBM India Research lab (IRL), we have developed a compact C++ based LVCSR acoustic modeling library IrlTK to estimate the maximum likelihood (ML) parameters of the hidden Markov model, Gaussian mixture model (HMM-GMM) based speech recognition systems using the standard ML re-estimation formulas[43]. IrlTK is capable of training state-of-the-art phonetic decision tree clustered triphone models. It builds a standard phonetic decision tree with the likelihood gain criterion for the node splitting. One tree is built for each of the state of the triphones that share the same middle phone and the node splitting stops when the gain in the likelihood of all the training samples under that node, falls below a certain threshold[31]. Finally all the triphones at the leaf nodes of a tree become the tied-states and their statistics are pooled together during the HMM-GMM parameter re-estimation. The training then proceeds with the gradual increase of the Gaussian components, re-estimating the HMM-GMM parameters in each iteration, to finally output tied-state triphone acoustic models. These models are evaluated through the following two decoders,

- A cross-word context dependent, beam-pruned, time-synchronous Viterbi search based *phonetic recognizer*. It uses a tied-state triphone acoustic model and a phonetic bigram language model and outputs a string of phonemes corresponding to an input utterance.
- A one-pass, beam-pruned, time-synchronous, Viterbi search based *continuous word recognizer (LVCSR)*[33] that uses word-internal triphone acoustic model. This decoder creates a dynamic search graph[30] and for simplicity uses a linear lexicon instead of a prefix tree based lexicon[29]. However, to reduce the redundant and expensive likelihood computation, it caches the evaluated likelihood of each active leaf (tied-state) at the current frame and if any other state that is tied to the same leaf becomes active, the cached likelihood value of the corresponding leaf is used instead of evaluating it again which otherwise would have led to redundant computation. Appropriate data-structures are used to achieve this effect. Token passing algorithm is used to keep the word histories as the linked-lists[34]. At each time frame, all those states whose accumulated likelihood upto that time frame fall below a certain threshold value (Beam) of the highest likelihood state, are pruned away from the Viterbi search space and thereby they do not propagate

new paths[30].

With these refinements, our LVCSR decoder with about $'2300'$ word vocabulary, a bigram language model and a $'20'$ component Gaussian mixture model with $'1800'$ leaves or tied-states (i.e. $1800 \times 20 = 36,000$ Gaussians) runs at about $6 \times RT$ (six times real time) on the SWB test-set. However, we note that the current instance of our LVCSR decoder is relatively simplified as compared to the mature and long-standing LVCSR decoders such as the HTK's `HDecode`[32] and other LVCSR decoders[29]. In particular, our present LVCSR decoder can only handle the word-internal triphone acoustic model and a bigram language model.

### C. MFCC and Fepstrum feature extraction

SWB speech data has a bandwidth of $8KHz$. We have used the standard 24 Mel-frequency filterbank covering $0-4KHz$to obtain the 13 dimensional Mel Frequency Cepstral Coefficients (MFCC) using a primary analysis window of size $30ms$ and a frame-shift of $10ms$. Their first and second order delta/derivative features were also computed to obtain a composite 39 dimensional MFCC feature vector which is used throughout this paper.

For the fepstrum feature extraction, we have used a $100ms$ long primary window to obtain a speech segment $x(n)$. At a sampling rate of $8KHz$, this corresponded to a $800$ sample speech segment $x(n)$. We then obtained its $800$ point discrete Fourier transform (DFT), denoted by $X(k)$. The analytic signal $s(n)$ of a real signal $x(n)$ is the inverse discrete Fourier transform (IDFT) of its positive frequency DFT (i.e. the DFT samples of X(k) that correspond to the angular frequency range $(-\pi, 0)$ are set to zero before taking the IDFT). Therefore, we set the last $400$ samples of the DFT (i.e $X(k)$) to zero as they corresponded to the frequency range $(-\pi, 0)$. This way we obtained the DFT $(S(k))$ of the analytic signal $s(n)$ directly. This was followed by the Mel-filterbank filtering in the frequency domain to separate the broadband analytic signal $s(n)$ into $24$ Mel-frequency sub-bands. The filtering operation was done in the frequency domain that simply involved multiplication of each of the Mel-filter's triangular frequency response with the DFT of the analytic signal $S(k)$. We performed IDFT on the filtered frequency response to finally obtain the Mel-filtered analytic signals for each of the $24$ Mel-bands. Thereafter, as per the equation (9), we took the logarithm of the absolute magnitude of each analytic signal in each band to obtain its AM signal. This schematic is also illustrated in Fig. (2). At this stage, the AM signal has the same sampling frequency as the original speech signal ($8KHz$). As can be noted in the Fig. 3, the AM signals are low modulation frequency signals. Therefore, we filtered the AM signals through a simple $40$ point moving average (MA) low-pass filter of cutoff-frequency $\approx 88Hz$[6] and then down-sampled them by a factor of $40$. We then retained its first $5$ DCT coefficients (Fepstrum) that roughly correspond to $[0, 25]Hz$ modulation frequency range. Fepstrum sub-vector from each band were concatenated together to form a super-vector of dimensionality $120$ ($5 \times 24$). As is well known, the adjacent frequency bands of the speech signal are correlated with each other. We therefore, performed a Principal Component Analysis (PCA) transform on this $120$ dimensional concatenated vector and retained only the top $60$ coefficients that corresponded to the top $60$ eigenvalues. These top $60$ eigenvalues accounted for almost $95.0\%$ of the total eigenvalues' mass. At this stage, this $60$ dim. feature vector can be concatenated with the standard $39$ dimensional MFCC feature vector or can be used of its own as a feature in a HMM-GMM based LVCSR system. We have evaluated fepstrum features in both the configurations and they are described next.

### D. Acoustic Model Training

The IrlTK library was used to train the maximum likelihood (ML) gender independent decision-tree tied-state triphone models. We have used ISIP's 2001 plain orthographic word level transcriptions of the SWB train set without any prior frame level segmentation in terms of phonemes or words. The IrlTK's training routine assumes that for a given time-frame, both the HMM (triphone) state and the Gaussian mixture component that emitted the feature vector, are hidden and hence performs full Forward-backward recursions[43] to collect the required sufficient statistics for estimating the parameters of the HMM-GMM model. While this provides considerable flexibility in the sense that we only needed plain word level orthographic transcriptions to train the models, it still required us to make subtle choices. For example, our word level orthographic transcriptions do not provide the definitive information whether there was an inter-word silence between a word pair. If we had a prior SWB acoustic model, we could have had used forced Viterbi alignment to get this information with some accuracy. However, as we were starting from scratch and we didn't have any prior SWB acoustic model, we decided to uniformly use a silence between every word pair. This transformed our word level transcriptions as illustrated below.

**Existing transcript** `sw202002.mfcc 5262 6164 UH HOWEVER I D[O]- I DO I DO LIKE A A LOT OF DIFFERENT FORMS OF MUSIC SO I SWITCH QUITE OFTEN UM I- I THINK I LIKE UH`

**Modified transcript** `sw202002.mfcc 5262 6164 SIL UH SIL HOWEVER SIL I SIL D[O]- SIL I SIL DO SIL I SIL DO SIL LIKE SIL A SIL A SIL LOT SIL OF SIL DIFFERENT SIL FORMS SIL OF SIL MUSIC SIL SO SIL I SIL SWITCH SIL QUITE SIL OFTEN SIL UM SIL I- SIL I SIL THINK SIL I SIL LIKE SIL UH SIL`

While we realize that this may be erroneous at certain word-pairs where there was no inter-word silence, we still decided to take this route as it was a uniform decision[7]. Further, this implicitly made our subsequent triphone transcriptions (obtained by replacing each word by its pronunciation and

[6] A $'M'$ point moving average filter's frequency response has the first null point at $f_{NULL} = \frac{f_s}{M}$ and the $3db$ cutoff at $f_{cutoff} = \frac{0.443 * f_s}{M}$, where $f_s$ is the sampling frequency.

[7] This may have also helped as it would have partially modeled the ever-present short-pause between the word-pairs.

grouping adjacent three phones as a triphone) as the *word-internal* triphones. The first triphone of each word had a SIL phone as the left context phone and similarly the last triphone of each word had a SIL as the right context phone. And, this also resulted in considerably lower complexity in the design of our LVCSR decoder as it implicitly had to use only the *word-internal* triphones.

Three feature sets were computed,

- [MFCC:] $39$ dimensional MFCC (primary window $30ms$).
- [Fepstrum:] $60$ dimensional Fepstrum (primary window $100ms$).
- [Concat. Fepstrum +MFCC:] $60$ dimensional fepstrum feature concatenated with the $39$ dim MFCC to obtain a $99$ dimensional feature.

We first set out to train our baseline system using the $39$ dimensional MFCC features. We had about $12,000$ triphones in our train set. The phonetic decision-tree based state tying resulted in about $1,800$ tied-state models covering all the $12,000$ triphones. Gaussian mixture components were gradually incremented to obtain $20$ component GMM for each tied-state. The recognition performance started to saturate around $17-20$ Gaussian components and hence we decided to keep the acoustic models with $20$ components per state. This resulted in about $1800 \times 20 = 36,000$ physical unique Gaussians in our acoustic model. Similarly, acoustic models with approximately $1800$ tied-states and $20$ component GMM were also trained for the fepstrum and fepstrum+MFCC features.

### *E. Phoneme Recognition*

Our first experiment consisted of a continuous phoneme recognition task on the $1.5Hr$ test-set. This was motivated by the TIMIT phoneme recognition task[48] which is widely used by the speech researchers to test new frontends (features) and new acoustic modeling techniques. In fact, our initial evaluation of the fepstrum features was on the TIMIT task where it provided about $1.5-2.0\%$ absolute improvement when concatenated with the MFCC features over the MFCC-only baseline system[21]. However, TIMIT consists of only about $4Hrs$ of total speech data and hence our present SWB setup provided us a much larger train and test sets, in addition to the challenging spontaneous speech phoneme recognition task, to evaluate these features. The above trained triphone acoustic models and a phoneme bigram language model (trained on the phonetic transcriptions of the train-set) were used. The continuous phoneme recognizer was the IrlTK context dependent triphone decoder described in the sub-section IV-B. The word insertion penalty and the LM weight factor were tuned on the disjoint $100$ utterance dev-set. In Tab I, we present the phoneme recognition accuracies of all the three feature sets. The fepstrum-alone feature resulted in an accuracy of $49.1\%$ as compared to the MFCC's accuracy of $51.1\%$. MFCC features had a good time-localization of about $20-30ms$. Whereas, the fepstrum features were estimated over a $100ms$ long window and hence did not possess a good time-localization. However, they complemented each other nicely with the fepstrum feature providing the modulation spectrum of each of the Mel-band in the range $0-25Hz$ and the MFCC providing the short-term spectral envelope information. Their simple concatenation was able to harness this information and provided $2.5\%$ absolute improvement over the MFCC baseline system.

TABLE I
PHONEME RECOGNITION ACCURACY USING PHONEME BIGRAM LM ON THE $1.5$ HR TEST-SET WITH $13$ CONVERSATIONS SPANNING $900$ UTTERANCES.

| System | Corr % | Acc % | Del % | Ins % | Sub % |
|---|---|---|---|---|---|
| MFCC | 58.0 | 51.1 | 21.9 | 7.0 | 20.0 |
| FEP | 55.8 | 49.1 | 23.8 | 6.7 | 20.4 |
| **MFCC+FEP** | **60.3** | **53.6** | 20.8 | 6.8 | 18.8 |

### *F. Continuous Word Recognition on SWB*

While the phoneme recognition experiments do provide useful insights in the evaluation of the new features, usually the final goal is the continuous word recognition. Therefore, we designed another set of experiments to evaluate the fepstrum features for the SWB word recognition task using our LVCSR decoder which has been described in the sub-section IV-B. In these experiments, we have used simple bigram language models (bgLM) trained on the raw word transcriptions of the SWB corpus which pre-dominantly consists of incomplete words, stuttering, self-repairs and word-compounding etc. In particular, we did not perform any particular word tokenization/normalization procedure. We simply used the plain raw word level orthographic transcriptions ($\approx 1.0M$ words) in our train set to create a bigram LM covering a vocabulary of size $2,300$ words. These raw word transcriptions exhibited a highly un-grammatical structure due to the presence of the stutterings, self-repairs, word-compoundings, repetitions, vocalized noise and laughter. Further, the topics of the conversations were highly diverse. The bigram LM trained on about $1.0M$ bigrams of the train-set transcriptions, resulted in a perplexity of $PP = 1180.7$ on our test-set word-transcriptions. This meant that the average branching factor after every word was $1180$ words. This extremely high perplexity was due the above mentioned factors and it clearly indicated the inadequacy of this LM for the decoding. As the typical perplexity range of the LMs used in the LVCSR is $PP = \{30, 100\}$, we decided to create a "test-set" matched LM with low perplexity. To this purpose, we interpolated the train-set bigrams ($\approx 1.0M$)and the test-set bigrams ($16,000$) with varying weights. We trained two bgLMs covering a vocabulary of size $2,300$ words.

- bgLM1: Used $1$ million train set bigrams with $weight = 1$ and $16,000$ test-set bigrams with $weight = 50$.
- bgLM2: Used only the $16,000$ test-set bigrams.

These LMs that are matched to the "test-set" conditions, resulted in a perplexity of $29.95$ and $19.12$ respectively. This meant that these were very good bigram LMs to perform the decoding of the test-set[8]. While, it would have been

[8] As our research focus was not exactly in training high-quality LMs from the unprocessed spontaneous speech word transcriptions, which remains an extensive research area in its own right, we decided to perform our recognition experiments using these "test-set" matched LMs.

ideal to obtain a LM that would not have seen the test-set bigrams, we believe that it was not a major limitation for our experiments due to the following two reasons. First, our acoustic models were indeed trained on the train-set which was completely disjoint with the test set and the fepstrum features, which we were evaluating, could have made their contribution only through the acoustic likelihoods. Second, the “test-set” matched LM was used for both the MFCC baseline system and the proposed fepstrum+MFCC system. Therefore, it would have helped both the systems during the decoding, in the similar ways. Therefore, any recognition accuracy improvement of one system over the another would have been most likely due to the differences in the features (front-ends) of the two systems.

In Tab. II, we show the recognition performance on the $1.5Hr$ test-set using the bgLM1. The word insertion penalty and the language model weight factor used in the decoder were tuned on the $100$ utterance dev-set which was disjoint from the test-set. We note that the MFCC system provided a word recognition accuracy of $45.2\%$ while the fepstrum feature alone provided an accuracy of $40.8\%$. MFCC features have good localization property as they were computed over a short-term window of $20-30ms$, while fepstrum features were computed over $100ms$. We, then concatenated the two features and trained an acoustic model. The composite feature benefited from the short term spectral envelope information in the MFCC and the AM modulation spectrum information in the fepstrum, thereby, resulting in an accuracy of $47.8\%$ – an improvement of $2.6\%$ absolute over the MFCC feature. Here, we also note that our acoustic model is a relatively basic acoustic model with ML word-internal triphones and without VTLN[41] and MLLR[42] adaptations. Further, we have used a simple one-pass decoder with just a bigram LM. Therefore, our recognition accuracies are not in the $60-65\%$ range as have been achieved by the long-standing speech groups[26], [27], [28] with multiple adaptations (VTLN, MLLR) and several pass decoding strategies with $3^{rd}-4^{th}$ order LMs. It is widely acknowledged that the LVCSR system design and build has now reached an extremely high-level of sophistication and complexity. On one hand, it reflects the maturity of the area and the impressive gains that have been achieved through the interplay of these multiple techniques. On the other hand, it has also partially impeded the rapid evaluation of new techniques on the LVCSR tasks due to the sheer complexity of developing and building these systems from the ground up. In these experiments, while we have not been able to implement the VTLN and MLLR adaptations, we have still endeavored to bring them as close as possible to the full-fledged LVCSR experiments. Therefore, our results are similar in spirit to the SWB results reported in [38], [39], [40] where the authors have also highlighted the ever-increasing trade-off between the development of the novel acoustic-modeling techniques, features and dealing with the LVCSR system development complexity.

TABLE II
WORD RECOGNITION ACCURACY USING BGLM1 ON THE $1.5$ HR TEST-SET WITH $13$ CONVERSATIONS SPANNING $900$ UTTERANCES.

| System | Corr % | Acc % | Del % | Ins % | Sub % |
|---|---|---|---|---|---|
| MFCC | 51.0 | 45.2 | 16.1 | 5.8 | 32.9 |
| FEP | 45.6 | 40.8 | 18.7 | 4.9 | 35.6 |
| **MFCC+FEP** | **53.2** | **47.8** | 15.3 | 5.4 | 31.5 |

Similar recognition experiments were also performed with the bgLM2. As, it was a perfect “test-set” conditions matched LM, it provided us a flavor of the accuracies that can be achieved with the low perplexity and possibly higher-order LMs[9] using the ML word-internal triphone models. The recognition accuracies are presented in Tab. III. As expected, we note that the recognition accuracies of all the features improved as compared to the use of bgLM1 in Tab. II. More importantly, the concatenation of the Fepstrum and MFCC feature provided an accuracy improvement of $3.5\%$ absolute over the MFCC feature. In this case, the Fepstrum+MFCC feature had an accuracy of $57.0\%$ as compared to the MFCC’s accuracy of $53.5$

TABLE III
WORD RECOGNITION ACCURACY USING BGLM2 ON THE $1.5$ HR TEST-SET WITH $13$ CONVERSATIONS SPANNING OVER $900$ UTTERANCES.

| System | Corr % | Acc % | Del % | Ins % | Sub % |
|---|---|---|---|---|---|
| MFCC | 60.1 | 53.5 | 12.8 | 6.6 | 27.1 |
| FEP | 55.7 | 50.1 | 15.2 | 5.7 | 29.1 |
| **MFCC+FEP** | **62.8** | **57.0** | 12.5 | 5.7 | 24.7 |

In all the above experiments, the increased performance of the concatenated Fepstrum and MFCC feature is due to the complimentary information present in these two feature-sets. While, the MFCC provided short term spectral envelope information with a good time-localization of $20-30ms$, the fepstrum provided the spectrum of the amplitude modulation (AM) signal in each Mel-band, estimated in a principled way. Fepstrum features were estimated over a window of size $100ms$, and covered a broad $0-25Hz$ range of the modulation frequency.

### *G. Qualitative comparison with the MLP transformed features*

In the recent years, multi-layer perceptron (MLP) transformed features, which in concatenation with the regular MFCC features, are input to the second-stage HMM-GMM systems, have shown improvements over the simple MFCC inputted HMM-GMM systems[16], [17], [22]. These improvements are due to two effects. First, the complementary information present in the new features that are used to train the first-stage MLP, gets presented in concatenation with the MFCC, to the training of the second-stage HMM-GMM system. This way, the second-stage HMM-GMM system can do some soft error-correction, during its own training, of the learned MLP outputs[10]. Second, MLP being a discriminative classifier, non-linearly transforms the features to increase the discrimination between the different phonemes. It is the combined effect and interplay of these phenomena that provides

[9] trained after word-tokenization, handling word-compounding, stutters, self-repairs on large text corpora that may be well representative of the test-conditions

[10] for details, please see [22]

the improvement. However, the use of the MLP transformed features[16], [17], [22] is not as straight-forward as the training of the HMM-GMM systems. MLP training requires frame level labels for the entire training set which becomes a limitation for the larger train sets which are not labeled at the frame level. Several researchers have proposed using a Viterbi forced alignment of the training set to obtain the frame level labeling of the train-set which pre-requisites an existing acoustic model for that speech data-set. In constrast to this, HMM-GMM system do not require the frame level labeling and can be trained with plain word level transcriptions. Given the word-level transcriptions, the Baum-Welch (EM) algorithm finds the state occupation probabilities required to train the HMM-GMM system[43]. This offers greater flexibility, especially in training acoustic models for the new acoustic environments, data-sets and languages. Additionally HMM-GMM systems are easily amenable to the various adaptation algorithms such as the MLLR[42] and VTLN[41]. Therefore, in this background, we believe that the proposed setup where the concatenated (fepstrum+MFCC) HMM-GMM system provides an absolute improvement of 2.5% over the MFCC HMM-GMM baseline may be a useful result as it provides the improvement while retaining the relative simplicity of the parameter training and adaptability of the HMM-GMM systems. Further, if the frame level labeling is available, fepstrum features too, can be MLP transformed and used in concatenation with the MFCC to train the subsequent HMM-GMM system. This was evaluated in [22], where the MLP transformed fepstrum features (in concatenation with the MFCC) resulted in a phoneme recognition accuracy of 76.6% and 77.6% on the TIMIT core test-set and complete test-set respectively, making them one of the best reported phoneme recognition accuracies[11] on the TIMIT task.

## V. Summary

In this paper, we presented a principled *non-parametric* technique for the AM-FM decomposition of the speech signals using the analytic signals in the time domain. We have shown that the previous techniques in the literature[14], [15], [16], [18], [23], [25], that extract the speech modulation spectrum by computing a discrete Fourier transform (DFT)/DCT of the Mel or critical band spectral energy trajectories (where each sample of the trajectory has been obtained through a short-term spectrum over $20-30ms$ long windows, followed by Mel-filtering) have two severe limitations. First, they implicitly assume that within each Mel or critical band, the AM signal remains constant throughout the duration of the window length that is typically $20-30ms$ long. Second, instead of modeling the continuously and slowly changing amplitude modulation signal in each band, they mostly model the spurious and abrupt modulation frequency changes that occur due to the subsequent frame shifting of $10ms$.

These features were implemented and evaluated in the context of the LVCSR experiments on a $72Hr$ portion of the Switchboard corpus using our recently developed LVCSR library IrlTK. A simple concatenation of the fepstrum and MFCC features benefited from the MFCC's short term spectral information and the fepstrum's complementary AM modulation spectrum information. This resulted in a phoneme recognition accuracy improvement of absolute $2.5\%$ over the MFCC features on a $1.5Hr$ SWB test-set. We further performed word recognition experiments on the same test-set and showed that the fepstrum+MFCC concatenated feature provided an absolute improvement of $2.6-3.5\%$ over the MFCC system. Most importantly, this accuracy improvement was obtained by a simple concatenation of the fepstrum features with the MFCC, without requiring the subsequent HMM-GMM recognizer architecture to be changed at all. Therefore, in a certain way, it may offer greater flexibility as compared to the MLP transformed feature extraction techniques[16], [17], [22]. This is due to the fact that MLP training requires a frame-labeled training data. Now, the frame-labeled training data can be obtained if a prior acoustic model exists for the training-data, and by performing a forced Viterbi alignment. However, if the frame-level segmentation is not highly accurate[12], the accuracy improvement may not be possible.

The proposed fepstrum feature, which only requires a simple concatenation with the MFCC feature, and can be used conventionally, with the plain word level orthographic transcriptions may offer a simpler and useful alternative while providing an accuracy improvement of $2.5-3.5\%$ absolute over the MFCC features. Therefore, they can be integrated and used with the existing HMM-GMM LVCSR recognition systems without much overhead. Indeed, if the frame level labeling of the train-set is available, fepstrum features too can be used to obtain the MLP transformed fepstrum feature, which then, can be inputed to the subsequent HMM-GMM system. This configuration has been evaluated on the TIMIT dataset in [22] resulting in positive gains.

Our future work will focus on fepstrum evaluation along with the MLLR and VTLN adaptations on the SWB word recognition task. We will further develop our LVCSR decoder enabling it to use higher-order LMs.

## References

[1] R. Kumaresan and A. Rao, " Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech appications, " J. Acoust. Soc. Am. 105(3), March 1999.

[2] R. Kumaresan, "An Inverse signal Approach to Computing the envelope of a real valued signal, " IEEE Signal Processing Letters, Vol. 5, No.10, October 1998.

[3] R. Kumaresan, "On minimum/maximum/all-pass decompositions in time and frequency domains, " IEEE Trans. Signal Processing, vol. 48, no.10, pp.2973-2976, Oct. 2000.

[4] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception, " Journal Acoustical Society of America, vol. 95, no. 5, pp. 2670-2680, 1994.

[5] H. Dudley, "Remaking Speech, " Journal Acoustical Society of America, vol. 11, no. 2, pp. 169-177, 1939

[6] H. Dudley, "The carrier nature of speech , " The Bell Systems Technical Journal, vol. 19, no. 4, pp. 495-515, Oct. 1940.

[7] O. Ghitza, "On the upper cutoff frequency of auditory critical-band envelope detectors in the context of speech perception, " Journal Acoustical Society of America, vol. 110, no. 3, pp. 1628-1640, 2001.

[8] S. Schimmel and L. Atlas, "Coherent Envelope Detection for Modulation Filtering of Speech, " Proc. of ICASSP 2005, Philadelphia, USA.

[11] including the deletion, insertion and substitution errors

[12] which in turn depends on the accuracy of the acoustic model used to obtain the Viterbi forced-alignment

[9] L. Atlas, and C. Janssen, “Coherent modulation spectral filtering for single-channel music source separation, ” Proc. of IEEE ICASSP 2005, Philadelphia, USA.

[10] S. Sukittanon, L. Atlas, J. W. Pitton and K. Filali, “ Improved Modulation Spectrum through Multi-scale Modulation Frequency Decomposition, ” Proc. of ICASSP 2005, Philadelphia, USA.

[11] S. M. Schimmel, “Theory of modulation frequency analysis and modulation filtering with applications in hearing devices, ” PhD Dissertation, Univ of Washington 2007.

[12] M. N. Murthi and B. D. Rao, “All-pole modeling of speech based on the minimum variance distortion-less response spectrum, ” IEEE Trans on Speech and Audio Processing, Vol.8, No. 3, May 2000.

[13] A. Potamianos and P. Maragos, “Time-Frequency Distributions for Automatic Speech Recognition,” In IEEE Tran. on Speech and Audio Prcoessing, Vol. 9, March 2001.

[14] N. Kanedera, H. Hermansky and T. Arai, “Desired Characteristics Of Modulation Spectrum For Robust Automatic Speech Recognition, ” In the Proc. of IEEE-ICASSP, 1998.

[15] B.E.D. Kingsbury, N. Morgan and S. Greenberg, “ Robust speech recognition using the modulation spectrogram,” Speech Communication, vol. 25, Nos. 1-3, August 1998.

[16] H. Hermansky, “ TRAP-TANDEM: Data driven extraction of the features from speech, ” In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA.

[17] Qifeng Zhu, Barry Chen, Nelson Morgan and Andreas Stolcke, ”On using MLP features in LVCSR,”, In the proc. of ICSLP 2004.

[18] V. Tyagi, I McCowan, H. Bourlard, H. Misra, “ Mel-Cepstrum Modulation Spectrum (MCMS) features for Robust ASR, ” In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA.

[19] V. Tyagi and C. Wellekens, “Fepstrum Representation of Speech Signal,” In the Proc. of IEEE ASRU 2005, Cancun, Mexico.

[20] V. Tyagi and C. Wellekens, “Fepstrum and carrier signal decomposition of speech signals through homomorphic filtering, ” In the Proc. of IEEE ICASSP, Toulouse, France, 2006.

[21] V. Tyagi, ”Fepstrum: An improved modualtion spectrum for ASR,”, In the Proc. of Interspeech 2007, Antwerp, Belgium.

[22] V. Tyagi, “Tandem Processing of Fepstrum features, ” In Proc. of Interspeech 2008, Brisbane.

[23] M. Athineos and D. Ellis, “Frequency domain linear prediction for temporal features, ” In the Proc. of IEEE ASRU 2003, St. Thomas, Virgin Islands, USA.

[24] M. Athineos and D. Ellis, “Autoregressive modeling of temporal envelopes, ” IEEE Trans. on Signal Processing, Vol. 55, no. 11, November 2007.

[25] Q. Zhu and A. Alwan, “AM-Demodualtion of speech spectra and its application to noise robust speech recognition,” Proc. ICSLP, Vol. 1, pp. 341-344, 2000.

[26] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V.R. Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, J. Zheng, “The SRI March 2000 Hub-5 Conversational Speech Transciption System, ” In the Proc. of NIST Speech Workshop, 2000.

[27] T. Hain, P. C. Woodland, T. R. Niesler, and E.W.D. Whittaker, “The 1998 HTK System for Transcription of Conversational Telephone Speech, ” In Proc. of IEEE ICASSP 1999.

[28] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. saon, G. Zweig, “The IBM 2004 Conversational Telephony system for rich transcription, ” In Proc of IEEE ICASSP 2005.

[29] H. Ney and S. Ortmanns, “Progress in Dynamic Programming Search for LVCSR, ” In Proc. of IEEE, Vol. 88, No. 8, Aug 2000.

[30] S. J. Young, “ A Review of the large vocabulary continuous speech recognition, ” IEEE Signal Processing Magazine, September 1996.

[31] S. J. Young, J. J. Odell and P. C. Woodland, “Tree based state tying for high accuracy acoustic modeling, ” Proc. of Human Language Technologies Workshop, pp. 307-312, Plainsboro, NJ, Morgan Kaufman Publishers Inc., March 1994.

[32] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University, 1995

[33] J. J. Odell, V. Valtchev, P. C. Woodland and S. Young, “A one pass decoder design for large vocabulary speech recognition, ” Proc. of Human Language Technologies Workshop, pp. 405-410, Plainsboro, NJ, Morgan Kaufman Publishers Inc., March 1994.

[34] S. J. Young, N. H. Russell, J.H.S. Thornton, “Token passing: A simple conceptual model for connected speech recognition systems, ” Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dept. 1989.

[35] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development, ” In Proc. of IEEE ICASSP, San Francisco, CA, USA, 1992.

[36] S. Greenberg, “The Switchboard Transcription Project, ” Technical Report of the 1996 LVCSR Summer Research Workshop, John Hopkins University, Baltimore, MD, USA, 1997.

[37] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, J. Picone, “Resegmentation of Switchboard”, In the Proc. of ICSLP, 1998.

[38] H. Jiang, and L. Deng, “A Robust Compensation Strategy for Extraneous Acoustic Variations in Spontaneous Speech Recognition, ” IEEE Trans. on Speech and Audio Processing, Vol. 10, No. 1, Jan. 2002.

[39] S. King, C. Bartels, and J. Bilmes, “SVitchboard 1: Small Vocabulary Tasks from Switchboard 1, ”, In Proc. of Interspeech 2005.

[40] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, K. Livescu, “An Articulatory feature based Approach and Factored Observation Modeling, ”, In Proc. of IEEE ICASSP 2007.

[41] L. Lee and R. Rose, “Speaker Normalization Using Efficient Frequency Warping Procedures, ” In Proc of IEEE ICASSP, Atlanta, GA, May 1996.

[42] C. J. Leggetter and P. C. Woodand, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models, ” Computer Speech and Language, Vol. 9, pp. 171-186, 1995.

[43] J. A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, ” ICSI and Computer Science Division, Dept. EECS, U.C. Berkeley, TR-97-021, April 1998.

[44] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear Prediction of the speech wave, ” J. Acoust. Soc. of America, Vo. 50, pp.637-655, Aug. 1971.

[45] J. Makhoul, “Linear Prediction: A Tutorial Review, ” In the Proc. of IEEE, vol.63, No.4, April 1975.

[46] S. Haykin, “Communication Systems, ” 3rd ed., pages 79-95, John Wiley & Sons, New York, 1994.

[47] A. V. Oppenheim and R. W. Schafer, “ Discrete-time signal Processing, ” Prentice-Hall, Inc., 1989, Englewood Cliffs, New Jersey.

[48] J.S. Garofolo, et al, “TIMIT Acoustic-Phonetic Continuous Speech Corpus, ” Linguistic Data Consortium, Philadelphia, 1993.