FEPSTRUM AND CARRIER SIGNAL DECOMPOSITION OF SPEECH SIGNALS THROUGH HOMOMORPHIC FILTERING

Vivek Tyagi and Christian Wellekens

Institute Eurecom
B.P 193 -06904 Sophia Antipolis, France
{Vivek.Tyagi, Christian.Wellekens}@eurecom.fr

ABSTRACT

Amplitude Modulation(AM) and frequency modulation(FM) have been well defined and studied in the context of communications systems[10]. Borrowing upon these ideas, several researchers have applied AM-FM[6, 7, 8, 9] modeling for speech signals with mixed results. These techniques have varied in their definition and consequently the demodulation methods used therein. In this paper, we carefully define AM and FM signals in the context of ASR. We show that for a theoretically meaningful estimation of the AM signal, it is necessary to decompose the speech signal into several narrow spectral bands as opposed to the previous use of the speech modulation spectrum[6, 7, 8, 9], which was derived by decomposing the speech signal into increasingly wider spectral bands (such as critical, Bark or Mel). Due to the Hilbert relationships, the AM signal induces a component in the FM signal which is fully determinable from the AM signal[1, 3]. We present a novel homomorphic filtering technique to extract the leftover FM signal after suppressing the redundant part of the FM signal. The estimated AM message signals are downsampled and their lower DCT coefficients are retained as speech features. These features carry information that is complementary to the MFCCs. A Tandem[4] combination of these two features is shown to improve recognition accuracy.

1. INTRODUCTION

In past several years, significant efforts have been made to develop new speech signal representations which can better describe the non-stationarity (spectral dynamics) inherent in the speech signal. Some representative examples are temporal patterns (TRAPS) features[4, 7] and the several modulation spectrum related techniques[6, 7, 8, 9]. In TRAPS technique, temporal trajectories of spectral energies in individual critical bands over windows as long as one second are used as features for pattern classification.

The notion of the amplitude modulation (AM) and the frequency modulation (FM) were initially developed for the communication signals[10]. In theory, the AM signal modulates a narrow-band carrier signal (specifically, a monochromatic sinusoidal signal). Therefore to be able to extract the AM signals of a wide-band signal such as speech (typically 4KHz), it is necessary to decompose the speech signal into narrow spectral bands. We follow this approach in this paper as opposed to the previous use of the speech modulation spectrum [6, 7, 8, 9] which was derived by decomposing the speech signal into increasingly wider spectral bands (such as critical, Bark or Mel). Similar arguments from the modulation filtering point of view, were presented by Schimmel and Atlas[2]. In their experiment, they consider a wide-band filtered speech signal $x(t) = a(t)c(t)$, where $a(t)$ is the AM signal and $c(t)$ is the broad-band carrier signal. Then, they perform a low-pass modulation filtering of the AM signal $a(t)$ to obtain $a_{LP}(t)$. The low-pass filtered AM signal $a_{LP}(t)$ is then multiplied with the original carrier $c(t)$ to obtain a new signal $\tilde{x}(t)$. They show that the acoustic bandwidth of $\tilde{x}(t)$ is not necessarily less than that of the original signal $x(t)$. This unexpected result is a consequence of the signal decomposition into wide spectral bands that results in a broad-band carrier[2]. We realise that this is not only a serious problem for modulation filtering[2], but also for modulation spectrum analysis (which is used as feature vector for ASR and is the topic of this paper).

Over the past few decades, pole-zero transfer functions that are used for modeling the frequency response of a signal, have been well studied and understood [5]. In this work we will denote them by "F-PZ". Lately, Kumaresan et al.[1] have proposed to model analytic signals[10] using pole-zero models in the temporal domain (denoted by T-PZ to distinguish them from the F-PZ). Along similar lines, Athineos et. al.[7] have used the dual of the linear prediction in the frequency domain to improve upon the TRAP features.

In [3], we have stated and proved several interesting time-frequency dualities for the analytic signals. These properties form the basis to develop "meaningful" AM-FM decomposition of the speech signal. There are two main contributions of this paper. Firstly, we develop a theoretically consistent AM signal analysis technique as compared to the previous ones[6, 8, 9]. We show that a "meaningful" AM signal estimation is possible only if we decompose the speech analytic signal into several narrow-band filters which results in narrow-band carrier signals. Secondly, we present a novel technique to extract the unique FM signal via homomorphic filtering route. We use the lower modulation frequency spectrum of the downsampled AM signal, as a feature vector (termed FEPSTRUM). The Fepstrum provides complementary information to the MFCC features and a Tandem[4] combination of the two features provides a significant ASR accuracy improvement over several other features.

This paper is divided into four sections. In Section 2, we describe the dual properties of the pole-zero models and the associated notation. In Section 3, the FM signal extraction is presented. Through examples, we show how the bandwidth of the analysis filter influences the estimated AM signal. In Section 4, experimental results are described followed by a conclusion in Section 5.

2. POLE-ZERO MODELS IN THE TEMPORAL DOMAIN

We recall that given a real periodic[1] signal $x(t)$ with period T seconds, its analytic version $s(t)$ is given by,

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

where $\hat{x}(t)$ denotes the Hilbert transform of $x(t)$. It has been shown in [3], that any arbitrary analytic signal can be expressed as a ratio of poles and zeros (T-PZ).

$$s(t) = a_0 e^{j\omega_t t} \frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \prod_{l=1}^{Q}(1 - q_l e^{j\Omega t}) \tag{2}$$

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[3].

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 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 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 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,

$$s(t) = a_0 e^{j\omega_t t} \underbrace{\frac{\prod_{i=1}^{P}(1 - p_i e^{j\Omega t})}{\prod_{i=1}^{U}(1 - u_i e^{j\Omega t})} \prod_{i=1}^{Q}(1 - 1/q_i^* e^{j\Omega t})}_{\text{T-MinP}} \times \prod_{i=1}^{Q}(-q_i^*) \underbrace{\prod_{i=1}^{Q} \frac{(e^{-j\Omega t} - q_i)}{(1 - q_i^* e^{-j\Omega t})}}_{\text{T-AllP}} \tag{3}$$

Our decomposition technique is based on the following lemma whose proof can be found in our previous work[3].

Lemma 1 *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* $\log|s(t)|$.

[1]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.

Therefore, using Lemma (1), $s(t)$ can be expressed as follows,

$$s(t) = \underbrace{a_0 \prod_{i=1}^{Q}(-q_i^*)}_{A_c} \underbrace{e^{\alpha(t)+j\hat{\alpha}(t)}}_{\text{T-MinP}} \underbrace{e^{j\gamma(t)}}_{\text{T-AllP}} \tag{4}$$

where A_c is a constant, $\alpha(t)$ is the logarithm 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. As $\hat{\alpha}(t)$ can be determined from the log AM signal $\alpha(t)$ [2], 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.

The next step is to develop algorithms that can automatically achieve the decomposition as in (4). 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], 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.

It is well know that the LP technique overestimates the peaks and poorly models the valley. 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 (4), 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.

3. CARRIER SIGNAL (FM) EXTRACTION

Fig.1 illustrates the FM signal extraction through homomorphic filtering. Consider a narrow band analytic signal $s_b(t), t \in [0, N-1]$. Our objective is to represent $s_b(t)$ as a product of a T-MinP signal and a T-AllP signal as done in (4). The phase of the T-AllP signal is the FM signal of interest. Let $realFepstrum(k) = FFT\{log|s_b(t)|\}$. In fact $realFepstrum(k)$ is the dual of the well known quantity i.e. real cepstrum. The causality conversion block in Fig.1 performs the following operation.

$$\begin{aligned} MinpFepstrum[k] &= 2 \times realFepstrum[k] \\ &\quad k \in [1, N/2 - 1] \\ MinpFepstrum[k] &= realFepstrum[k] \\ &\quad k = 0, N/2 \\ MinpFepstrum[k] &= 0 \\ &\quad k \in [N/2 + 1, N - 1] \end{aligned} \tag{5}$$

As $MinpFepstrum[k]$ is a causal sequence, its IFFT, $log(sMinp(t)) = \beta(t) + j\hat{\beta}(t)$ is an analytic signal. In light of lemma (1), it can be seen that, $sMinp(t) = e^{\beta(t)+j\hat{\beta}(t)}$ is the T-MinP signal in (4). Moreover, according to (4), T-AllP signal $sAllP(t)$ that corresponds to the original signal $s_b(t)$ can be obtained as,

$$sAllP(t) = \frac{s_b(t)}{sMinp(t)} \tag{6}$$

The unique FM signal can be obtained from $sAllP(t)$ as the derivative of its unwrapped phase. We have been experimenting with the

[2]Due to the HT relationship between the two

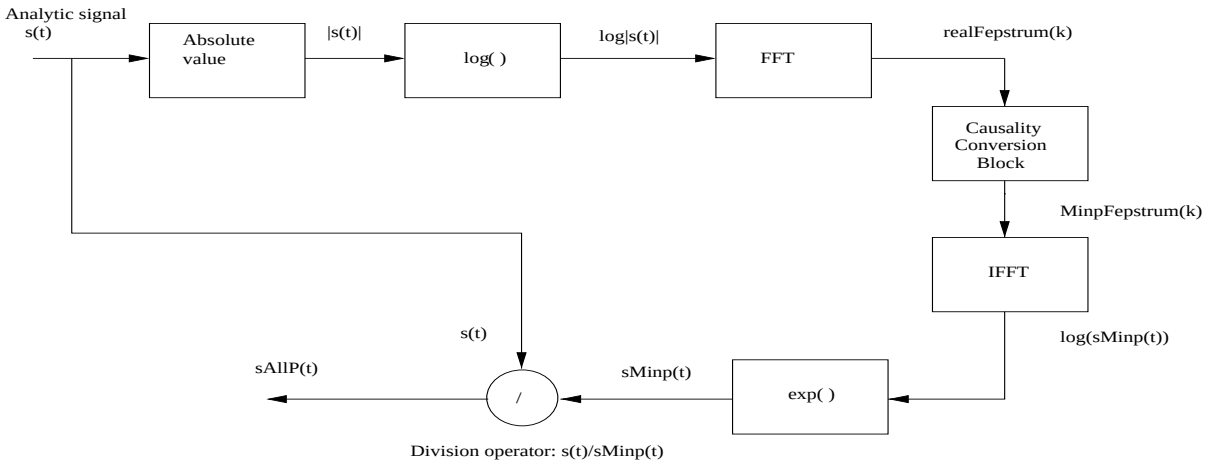

Fig. 1. Carrier signal decomposition via homomorphic filtering.

signal $sAllP(t)$ to gather an understanding of this new signal and the ways in which it can be used as a feature that is suitable for ASR. However, in this work, we have used only the AM modulation spectrum as a feature and the work on FM signal inclusion in under progress.

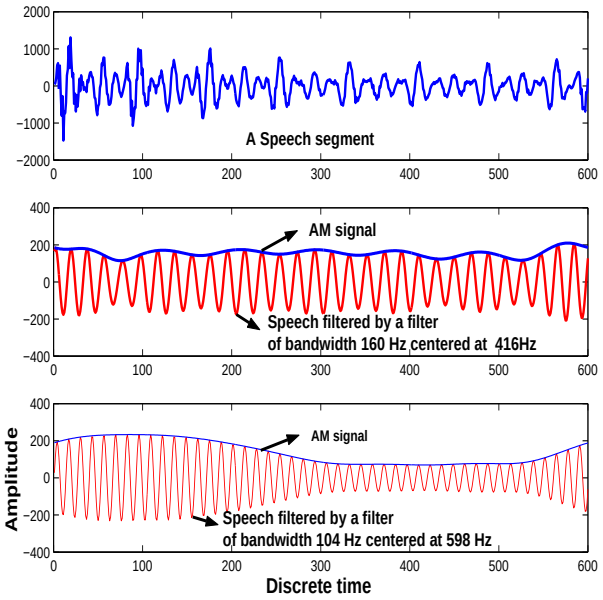

Fig. 2. The AM signal derived using narrow-band filters

Please refer to [3] to find a description of the AM signal extraction scheme. Fig.2 illustrates the case when we use narrow-band filters to decompose the speech analytic signal, followed by the AM signal estimation in each band. Second and third pane shows the narrow band-pass filtered speech signals and their corresponding AM signals. We note that these AM signals are low modulation frequency signals. The narrow band-pass filters used have band-widths 160 Hz and 104 Hz respectively. Fig.3 illustrates the case where a broad-band filter (bandwidth 533 Hz) has been used. For voiced speech, each pitch harmonic can be roughly

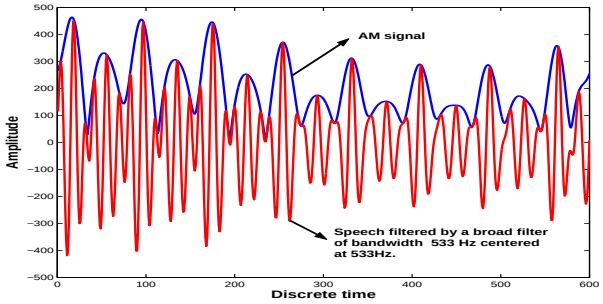

Fig. 3. The AM signal derived using broad-band filters

seen as a monochromatic sinusoidal carrier signal. The spectrum of the signal at the output of a broad-band filter will have several pitch harmonics in it and therefore will violate the condition of a narrow-band carrier signal. As can be noted in the Fig.3, the pitch component manifests itself as sharp spikes in the AM signal. Therefore a modulation spectrum of this AM signal will reflect the pitch frequency as well, which is undesirable in the context of a speaker independent ASR system. We present these arguments to justify our choice of the non-overlapping narrow-band filterbank instead of a critical, Bark or Mel-scale filterbank, in the Fepstrum estimation.

4. EXPERIMENTS AND RESULTS

In order to assess the effectiveness of the fepstrum features, speech recognition experiments were conducted on the OGI Numbers corpus [11]. It consists of spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words.[3]. We used 20 linearly spaced, non-overlapping rectangular filters to decompose the speech analytic signal into narrow-band signals of bandwidth 200Hz each. The AM signal is obtained as the logarithm of the absolute magnitude of the narrow-band filter output. At this stage, the AM signal has the same sampling frequency as the original speech signal ($8KHz$). As can be noted

[3]with confusable words like nine, ninety and nineteen.

in the Fig2, the AM signals are low modulation frequency signals. Therefore, we filter the AM signals through a low-pass filter of cutoff-frequency 200 Hz and then downsample them by a factor of 40. Long rectangular windows of size 85 ms were used to frame the narrow band-pass filtered analytic signals. This was done to ensure that we have sufficient number of samples after downsampling the AM signal. We chose a rectangular shape of the window to avoid any artificial tilt in the lower DCT coefficients. We then retain its first 5 DCT coefficients (Fepstrum) that correspond to $[0, 50]$ Hz. Fepstrum sub-vector from each band are concatenated together to form a vector of dimensionality 100 (5×20). We perform a KL transform on this vector, followed by dimensionality reduction to obtain a 60 dim. feature vector. These features are then fed to a trained multi-layer perceptron (MLP) to obtain phoneme aposteriors which are again KL transformed to obtain 27 dimensional Tandem-Fepstrum features[4]. Tandem[4] has been shown to be an effective technique for combining different kind of features. Fepstrum being a modulation spectrum carries information that is complementary to the usual spectral envelope based MFCC features. Therefore we have concatenated MFCC feature with the Fepstrum-Tandem features.

Mel-frequency cepstral coefficients (MFCC) and their temporal derivatives along with cepstral mean subtraction have been used as additional features. For comparison, four feature sets were generated:

1. [T-MFCC:] 27 dim. Tandem representation of MFCC + delta features.
2. [T-Fepstrum:] 27 dim. Tandem representation of Fepstrum features
3. [Concat. MFCC+ (T-MFCC):] (27+39) dim. feature vector which is a concatenation of the MFCC and Tandem-MFCC
4. [Concat. MFCC+ (T-FEPSTRUM):] (27+39) dim. feature vector which is a concatenation of the MFCC and Tandem-Fepstrum features.

All the above features were then used in a Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition system that was trained using public domain software HTK on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMMs with 3 emitting states per triphone and 12 mixtures per state. Table 1 indicates the performance of these feature sets. T-Fepstrum features have only the modulation frequency information and hence they perform slightly worse than the T-MFCC feature. However, as they carry complementary information, their concatenation (MFCC+T-FEPSTRUM) results in the lowest (4.1%) WER. For a fair comparison, we compare this to a concatenation of the MFCC+ T-MFCC features which has a WER of 4.6%. This is an encouraging result and we are further working on the feature represenations that will include the FM signal information in it. This may improve the results further.

Table 1. *Word error rate (WER) in clean conditions*

T-MFCC	5.2
T-FEPSTRUM	5.5
Concat. MFCC+ (T-MFCC)	4.6
Concat. MFCC+ (T-FEPSTRUM)	4.1

5. CONCLUSION

We have extended the work of Kumaresan[1] to develop a theoretically sound AM-FM decomposition technique suitable for ASR application. We point out to the deficiency in the previous use of the modulation spectrum that was caused due to the use of the broad-band filters. A novel homomorphic filtering based algorithm was presented to extract the FM signal of interest. Finally we present a suitable representation of the AM signal in form of the lower modulation frequencies of the downsampled AM signals in each band. A concatenation of the MFCCs with the Tandem-Fepstrum features achieves the lowest WER (4.1%). We are further working on the representations of the extracted FM signal that are suitable for ASR.

[4] each dimension corresponds to a monophone which are 27 in number

6. ACKNOWLEDGMENTS

This work was supported by European Commission 6th Framework Program project DIVINES under the contract number FP6-002034.

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

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

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

[5] 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.

[6] 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.

[7] M. Athineos, H. Hermansky, D. Ellis, "LP-TRAP: Linear predictive temporal patterns, ", Proc. of SAPA, Jeju, S. Korea, April 2004.

[8] 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.

[9] 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.

[10] S. Haykin, "Communication Systems, " 3rd ed., pages 79-95, John Wiley Sons, New York, 1994.

[11] R. A. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at CSLU," Proc. of ICSLP, Yokohama, Japan, 1994.