IDIAP RESEARCH REPORT

AM-FM DECOMPOSITION OF SPEECH SIGNAL: APPLICATIONS FOR SPEECH PRIVACY AND DIAGNOSIS

Petr Motlicek Hynek Hermansky
Srikanth Madikeri Amrutha Prasad
Sriram Ganapathy

Idiap-RR-01-2020

JANUARY 2020

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

AM-FM DECOMPOSITION OF SPEECH SIGNAL: APPLICATIONS FOR SPEECH PRIVACY AND DIAGNOSIS

Petr Motlicek[1], Hynek Hermansky[2], Srikanth Madikeri[1], Amrutha Prasad[1], Sriram Ganapathy[3]
Idiap Research Institute, Martigny, Switzerland[1]
The Johns Hopkins University, Baltimore, USA[2]
Indian Institute of Science Bangalore[3]

{petr.motlicek,srikanth.madikeri,amrutha.prasad}@idiap.ch, hynek@jhu.edu, sriramg@iisc.ac.in

Abstract: **Although current trends in speech processing consider deep learning through data-driven technologies, many potential applications exhibit lack of training or development data. Therefore, considerably light signal processing techniques are still of interest. This paper describes an efficient technique for decomposing the AM and FM components of the speech signal, which is not based on frame-by-frame short-time analysis of the signal. Instead, we estimate all-pole models of frequency-localized Hilbert envelopes of large segments of speech signal at different frequencies. The technique on decomposition of speech signal into AM and FM components appears to be of interest in voice studies benefiting from alleviation of the message-bearing components of speech (e.g. security oriented applications such as speaker recognition, or speech diagnosis often relying on spectra averaging to discard the content of the speech). Similarly, discarding speaker information while preserving the message in the speech is of interest for privacy-oriented applications. Experimental results on automatic speech and speaker recognition tasks clearly show that the AM component preserves the content (message) of the speech, while the FM component carries the information related to the speaker.**

Keywords: **AM, FM, Linear prediction, Automatic speech recognition, Speaker recognition**

I. INTRODUCTION

Dominant view of speech signal processing is still based on the linear model of speech production, where short segments of the signal (short enough so that the vocal tract does not significantly change within the segment) can be represented by short-time spectrum computed from these segments. The short-time spectrum consists of its spectral envelope (representing a linear filter emulating vocal tract transfer function at a given time instant) and its fine spectral structure. It is widely accepted that the spectral envelope mainly represents the phonetic value of the speech segment (i.e. message) and the fine structure represents the spectrum of the excitation source. Many speech-oriented applications would benefit from being able to reliably separate contributions of the signal excitation and of the filtering.

Typical conventional techniques, such as linear prediction (LP) [7], are based on the linear modeling and apply frame-by-frame inverse filtering of speech using estimates of spectral envelopes of short speech segments. In this paper, we abandon the notion of the short-time spectrum of speech. Instead, we (along with work of Dudley 1940 [5]) see the speech as an audible signal generated by voice source (frequency modulated component FM), which is modulated by inaudible and mostly invisible movements of the vocal tract (amplitude modulated component AM). The movements of the vocal tract carry a bulk of the message in speech, while the voice source makes these tract movements audible, allowing for the message to be perceived by a listener.

The paper describes an efficient technique for decomposing the AM and FM speech components, not based on frame-by-frame short-time analysis. Instead, we estimate all-pole models of frequency-localized Hilbert envelopes of large speech segments at different frequencies. This is done by applying the LP technique to short segments of a cosine transformed speech signal. Since each segment of the cosine transformed signal represents the individual frequency component of the original signal, the resulting all-pole models yield the frequency-localized Hilbert envelopes of the signal. Inverse cosine transforms of their LP residuals then yield frequency-localized FM components of the voice source signal. Summing all frequency-local FM estimates yields the FM voice signal with its message alleviated. When the audible AM component of the speech signal is desired, the frequency-localized all-pole models of Hilbert envelopes are used to compute frequency-localized modulated noise

components, which are summed to yield the AM signal component carrying the speech message.

II. AM-FM DECOMPOSITION

The concept of AM-FM decomposition is presented through frequency domain linear prediction (FDLP) - an efficient technique for autoregressive modelling of temporal envelopes of the signal [8]. FDLP proposes to model the speech in critical bands as a modulated signal with the AM component obtained using Hilbert envelope estimate and the FM component obtained from the Hilbert carrier. The sub-band temporal envelopes can then be estimated using FDLP. Unlike traditional temporal domain LP representing the envelope of the power spectrum of the signal [7], FDLP particularly exploits the prediction power of slowly varying long-term AM envelopes of speech signals in critical sub-bands. The final FDLP model provides smoothed, minimum phase representation of temporal rather than spectral envelopes.

The duality between time and frequency domains suggests that the power of autoregressive models can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples. Interestingly for FDLP, it has been analytically shown that the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal. At the same time it is known that the quadrature version of a real input signal and its Hilbert transform are identical for many modulated signals, known in practice. We can therefore presume that the Hilbert envelope approximates squared AM envelope of the signal. Thus, FDLP estimates the AM envelope of the signal and the FDLP residual contains the FM component of the signal. Acoustic signals in sub-bands are modulated signals and hence, FDLP can be used for AM-FM decomposition of sub-band signals.

III. DETAILED ANALYSIS

Source-filter linear model of speech production: Our current view of speech is dominated by the concept of the linear model of speech production (Chiba and Kajiyama 1942) [6], where the stationary source signal is filtered by the stationary filter. It assumes no interaction between the two components of this model (hence "linear"). This model is the basis for the LP speech analysis.

Carrier nature of speech: Before Chiba and Kajiyama, Homer Dudley (Dudley 1940) [4] published his concept of speech, where he suggested that for the human communication by speech, nature evolved a technique which is conceptually identical to the (then dominant) AM radio communication. In his concept, messages are carried in signal changes, reflected in slow movements of vocal tract. The movements are made audible by using them for modulating the audible voice carrier. The current paper follows this concept in the form of the FDLP.

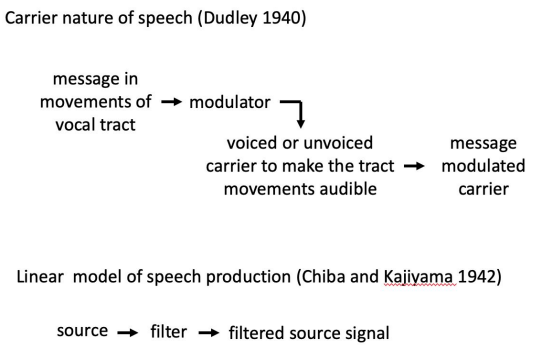

Fig. 1: Dudley's concept of speech [5] as a modulated carrier signal and the linear model of speech production.

Estimating components of models of speech: In deriving the speech messages, we are primarily interested in the vocal tract movements, i.e., in the modulation function. On the other hand, in many applications of voice technologies such as a speaker recognition, or voice pathologies, it is the carrier, which is of interest.

Conventional method of the carrier extraction is inverse filtering, where estimated spectral envelopes of short speech segments are used for design filters, which are then used for whitening the respective short segments of speech signals. A typical example of this technique is the LPC inverse filtering [7]. This in effect yields the modulating function, which is sampled at the frame-rate of the short-time analysis. Since the assumptions of stationarity and linearity are easily violated, an accurate estimation of the individual components of this model can be difficult [9].

We are following the original Dudley's concept, where estimated temporal envelopes of spectral trajectories of speech signals at different frequencies are used for alleviating message components in respective frequency bands. Estimating the modulating function was originally done by analog low-pass filtering of spectral energies in different frequency bands [4]. Here, we show that the concept of the all-pole modeling employed in the LP analysis can be successfully adopted for the estimation of spectral energy trajectories in different frequency bands.

IV. FDLP

The concept of the FDLP for modeling short segments of Hilbert envelopes was investigated in [11] and extended by modeling of Hilbert envelopes in narrow frequency bands in [12,8].

In FDLP, the LP prediction is applied to the cosine transform of the speech signal *s(t),* $t \in \langle 0, T \rangle$. One way to compute the cosine transform $q(t)$, $t \in \langle 0, T \rangle$ of a signal *s(t)* is through the Fourier transform of the signal $S_{sym}(t)$, $t \in \langle 0, 2T \rangle$, which is the even symmetrized *s(t), i.e.,* $q(\omega) = F[S_{sym}(t)]$. The $q(\omega)$ is a function of frequency and is real and even symmetric.

Being after the cosine transform in frequency domain allows for a selection of the frequency range to be further processed. The signal

$q_w(\omega) = q(\omega)\, w(\omega)$, where window

$w(\omega_0) = \{w_\omega \quad \omega = -\Delta\omega \le \omega_0 \le \Delta\omega\ 0,$ *otherwise.*

ω_0 indicates the center of the frequency band to be processed. The Fourier transform of $q_w(\omega)$, which is still real and causal, obeys the Krammers-Kroening relation $F\left[q_w(\omega)\right] = \left\{s_{\omega_0}(t) + H[s_{\omega_0}(t)]\right\}$.

The signal in a given frequency band, centered at ω_0, $s_{\omega_0}(t)$, now stands in place of the real part of the Fourier transform and its Hilbert transform takes place of its imaginary part. The instantaneous energy in the signal in a given frequency band (Hilbert envelope) $H_{\omega_0}(t) = s_{\omega_0}(t)^2 + H[s_{\omega_0}(t)]^2$ is an equivalent of the power spectrum $P(\omega)$ in the time-domain LP.

The autoregressive model computed from the cosine transform of the signal $q(\omega)$ obeys the equation

$$E_{\omega_0} = \frac{1}{2T} \int_{-T}^{T} \frac{H_{\omega_0}(t)}{\hat{H}_{\omega_0}(t)} dt,$$

where $\hat{H}_{\omega_0}(t)$ is the all pole autoregressive model of the Hilbert envelope $H_{\omega_0}(t)$ and E_{ω_0} is the error of the model fit in the frequency band centered at ω_0 over the time interval *T.* The form of the error equation implies a good fit of the spectrum of the autoregressive model $\hat{H}_{\omega_0}(t)$ to the peaks of the Hilbert envelope $H_{\omega_0}(t)$. Center of the frequency window $w(\omega_0)$ is typically gradually moved through the whole frequency range of the signal to be processed.

Re-synthesis from the FDLP: The FDLP model can be used to construct inverse filter for whitening the segment of the cosine transform. Whitened segment is inverse cosine filtered to represent the whitened signal in the respective band. Adding whitened signals from all frequency bands yields the carrier signal. Modulating white noise in the frequency bands by the estimated FDLP Hilbert envelopes yields whispered-like speech with the original speech message.

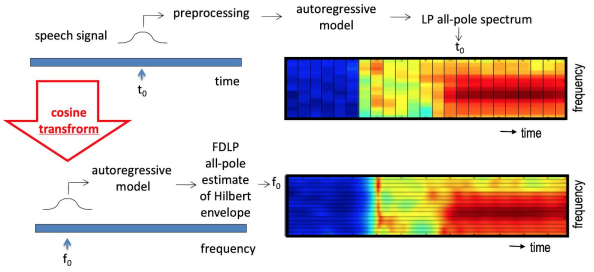

Fig. 2: The upper part of the picture shows the conventional LP as used in estimation of short-time spectral envelopes of short segments of speech centered at different times t0. The lower part shows the process of estimation of Hilbert envelopes in different frequency bands of speech signal centered at f0.

V. APPLICATIONS

The technique on decomposition of speech signal into AM and FM components appears to be of interest in voice studies, which would benefit from alleviation of the message-bearing components of speech (e.g. security oriented applications such as speaker recognition, or speech diagnosis often relying on spectra averaging to discard the content of the speech). In this paper, we empirically show that AM and FM components of the speech signal carry different types of information, AM related to the content and FM related to the speaker information, respectively.

V. EXPERIMENTS

We apply the AM-FM decomposition proposed in [10]. FDLP approach described in Section IV uses a simple window on top of cosine transformed (1000 ms long) speech segment to select a particular frequency band. Unlike previous, the following experiments apply slightly different FDLP version, available freely at Github[1]. First, the input speech is decomposed into 32 critically-sampled frequency sub-bands by using a conventional quadrature mirror filter (QMF) bank. FDLP is then applied on each sub-band to model the sub-band temporal envelopes (AM components). The LP residual represents the FM in the sub-band signal. These steps are reversed at the synthesis side, to reconstruct the signal back from QMF sub-band components.

Two sets of experiments are performed: automatic speech recognition (ASR), and speaker verification (SV) deployed on (i) original (fullband) speech, (ii), the speech reconstructed only from the AM sub-band components (i.e. envelope extracted using FDLP), and

[1] github.com/iiscleap/SignalAnalysisUsingAm-FM

Tab. 1: ASR and SV results measured in terms of word error rate (EER) and equal error rate (EER), respectively, on Librespeech corpus.

	ASR system	SV system
	WER [%]	EER [%]
Original speech	10.1	14.7
AM-only	14.9	26.5
FM-only	53.9	25

(iii) the speech reconstructed only from the FM sub-band components (i.e. carrier part alone). Subjective listening tests clearly show that the AM-only reconstructed signal sounds whispered. With the carrier part alone, the synthesized signal sounds message less.

Dataset and tool: For ASR and SV experiments, we use Librispeech corpus [3] which consists of read speech from audio books. We employ 100 hours for training (train-clean-100) and 5.4 hours for testing (test-clean). Kaldi toolkit [2] is used for building both ASR and SV.

ASR: the system is built around a conventional HMM-GMM framework. We use standard Kaldi (tri4) recipe comprising MFCC features projected by LDA+MLLT [1]. Roughly ~3.5K triphones and ~40k Gaussians are used to build HMM-GMM.

SV: Gaussian Mixture Models (GMMs) with 32 components are trained for each speaker in test set. Each GMM is built with the expectation-maximization algorithm to maximize the likelihood of the data [13]. Only 10s of speech data were used for both GMM development and testing. Cross-pair trials for SV experiments were generated and trials comparing the same audio are excluded. T-norm is applied on the test scores.

VI. DISCUSSIONS AND CONCLUSIONS

The paper discusses employment of AM-FM decomposition to efficiently alleviate message bearing components from the speech. The technology is demonstrated on ASR and SV tasks. As can be seen from Tab. 1, the speech signal reconstructed from AM components yields WER~14.9%, close to the performance of the original signal (WER~10.1%) on the standard ASR task. On the other side, the speech reconstructed from FM-only components largely increases WER (~53.9%). In the case of SV task, the obtained results are less obvious. Original speech still provides the best performance (EER~14.7%) as the SV engine also exploits the content to model the speaker. Nevertheless, the speech signal reconstructed from FM-only components still outperform AM-only speech (EER~25%) which clearly indicates that the speaker related information is preserved by the Hilbert carrier.

FDLP technique described in this paper, allowing to decompose the speech into AM and FM components, operates on large segments of signal at different frequencies. Empirically obtained results on automatic speech and speaker recognition tasks confirm our assumptions (determined by subjective listening) that the AM-FM decomposition can reliably separate the content and speaker related information from speech, which can be applied in various speech-oriented tasks.

REFERENCES

[1] S. Rath, et al, “Improved feature processing for Deep Neural Networks,” *Proc. of Interspeech 2013*.

[2] D. Povey, A. Ghoshal, et al., “The Kaldi Speech Recognition Toolkit,” *in Proc. of IEEE ASRU*, 2011.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” *in IEEE ICASSP*, 2015.

[4] H. W. Dudley, ``The vocoder," Bell Labs Rec., vol. 18, pp. 122-126, 1939.

[5] Dudley, H. (1940). The carrier nature of speech. *Bell System Technical Journal*, 19(4), pp. 495-515.

[6] T. Chiba, M. Kajiyama (1958), “The vowel: Its nature and structure” (Vol. 652), Tokyo: Phn. society of Japan.

[7] J. Makhoul (1975), “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, 63(4), pp. 561-580.

[8] M. Athineos, D. Ellis, "Frequency-domain linear prediction for temporal features", *Proc. IEEE ASRU Workshop*, pp. 261-266, December 2003.

[9] Alku, P. (1992). “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech communication*, 11(2-3), pp. 109-118.

[10] S. Ganapathy, P. Motlicek and H. Hermansky, "Autoregressive Models Of Amplitude Modulations In Audio Compression", *IEEE Transactions on Audio, Speech and Language Processing*, August 2010.

[11] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal", *IEEE Signal Proc. Letters*, vol. 5, no. 10, pp. 256-259, 1998.

[12] J. Herre, J. D. Johnston, (1996, November), “Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS),” *In Audio Engineering Society Convention 101.*

[13] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech communication* 17.1-2 (1995): pp. 91-108.