Time-Frequency Distributions for Automatic Speech Recognition

Alexandros Potamianos*, *Member IEEE* and Petros Maragos†, *Fellow IEEE*

* Bell Laboratories, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07074, U.S.A.

† Dept. of ECE, National Technical University of Athens, Zografou 15773, Athens, Greece.

May 14, 2008

Abstract

The use of general time-frequency distributions as features for automatic speech recognition (ASR) is discussed in the context of hidden Markov classifiers. Short-time averages of quadratic operators, e.g., energy spectrum, generalized first spectral moments, and short-time averages of the instantaneous frequency, are compared to the standard front end features, and applied to ASR. Theoretical and experimental results indicate a close relationship among these feature sets.

EDICS: 1-RECO, 1-ANLS

Some of this work was performed while the authors were with the School of E.C.E, Georgia Institute of Technology, Atlanta, GA 30332, USA. It was partially supported by the US National Science Foundation under Grants MIP-9396301 and MIP-9421677. P. Maragos' recent work is supported by the Greek G.S.R.T. program in Language Technology, Grant 98GT26.

1 Introduction

Time-frequency distributions and short-time averages of quadratic operators are very popular front-end features for automatic speech recognition (ASR). Indeed, the "standard" front-end feature set is the inverse cosine transformation of the short-time–frequency energy distribution. Despite the standardization of the ASR front-end, there has been a significant amount of research on using alternate time-frequency distributions as (possibly additional) ASR features. A good review of such efforts can be found in [7]. However, such efforts are often lacking in theoretical or experimental justification. In this paper, we attempt to outline the relationships between some popular alternative feature sets and the "standard" front-end features, and to present experimental ASR evidence that supports these claims. We hope that this study will help guide future ASR front-end research.

The following two types of non-parametric features are investigated in this paper: (i) short-time averages of quadratic operators, e.g., energy spectrum [8], (ii) generalized first spectral moments and weighted short-time averages of the instantaneous frequency. Note that the standard feature set is included in the first family of time-frequency distributions. Our goal is to show (both theoretically and experimentally) a close relationship among these feature sets and the standard feature set.

The organization of the paper is as follows: First, we introduce the energy operator and the energy spectrum, and compare it to other spectral envelope representations. In Section 3, short-time instantaneous frequency estimators are proposed in the context of the AM–FM modulation model, the sinusoidal model, and spectral estimation. The estimators are compared to the spectral envelope and their merits as ASR features are discussed. Finally, experimental ASR results are given in Section 4. The authors assume in the presentation some familiarity with the sinusoidal speech model [5], the AM–FM modulation model [3] and energy operators [2, 4].

2 Quadratic Operators and Energy Spectrum

The energy operator is defined for continuous-time signals $x(t)$ as

$$\Psi_c[x(t)] \stackrel{\triangle}{=} [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \qquad (1)$$

where $\dot{x} = dx/dt$. Its counterpart for discrete-time signals $x(n)$ is

$$\Psi_d[x(n)] \stackrel{\triangle}{=} x^2(n) - x(n-1)x(n+1) \qquad (2)$$

The nonlinear operators Ψ_c and Ψ_d were developed by Teager during his work on speech production modeling [11] and were first introduced systematically by Kaiser [2]. When Ψ_c is applied to signals produced by a simple harmonic oscillator, e.g. a mass-spring oscillator, it can track the oscillator's energy (per half unit mass), which is equal to the squared product of the oscillation amplitude and frequency; thus the term *energy operator*. The energy operator has been applied successfully to demodulation and has many attractive features such as simplicity, efficiency, and adaptability to instantaneous signal variations [3]. The attractive physical interpretation of the energy operator has led to its use as an ASR feature extractor in various forms, see for example [12, 13].

The *energy spectrum*, introduced in [8], is a general time-frequency distribution based on the energy operator. Assume that $x(n)$ is filtered by a bank of K bandpass filters centered at frequencies ω_k to obtain K band-passed signals: $x_k(n)$, $k = 1..K$. The following time and frequency relations hold

$$x_k(n) = x(n) * h_k(n) \leftrightarrow X_k(\omega) = X(\omega)H_k(\omega) \tag{3}$$

where $h_k(n)$ is the impulse response and $H_k(\omega)$ is the frequency response of the kth filter and n is the discrete-time sample index. The energy spectrum $ES(n,k)$ is defined as the short-time average of the energy operator applied to the family of band-passed signals $x_k(n)$, i.e,

$$ES(n,k) = \sum_{m=n}^{n+N-1} \Psi_d[x_k(m)] \tag{4}$$

where N is the length of the short-time averaging window (in samples). Using Parseval's relation one can show

$$\sum_{n=-\infty}^{\infty} x(n+a)x(n+b) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos[(a-b)\omega]|X(\omega)|^2 d\omega \tag{5}$$

assuming that $x(n)$ is real. Thus

$$\sum_{n=-\infty}^{\infty} \Psi_d[x(n)] = \frac{1}{2\pi}\int_{-\pi}^{\pi} [1-\cos(2\omega)]|X(\omega)|^2 d\omega. \tag{6}$$

Assuming that $x(n)$ is zero outside of the window $[n, n+N-1]$ the energy spectrum can be expressed as

$$ES(n,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} [1-\cos(2\omega)]|H_k(\omega)X(\omega)|^2 d\omega. \tag{7}$$

In Fig. 1, the time-domain implementation of a general filterbank-based ASR front-end is shown. Following the notation introduced above $x(n)$ is filtered by a bank of K filters. The feature set at time index n $\{TF(n,k) : \ k = 1..K\}$ is defined as the short-time average of the output of a quadratic operator $Q(.)$ applied to each one of the band-passed signals $x_k(n)$, i.e.,

$$TF(n,k) = \sum_{m=n}^{n+N-1} Q[x_k(m)]. \tag{8}$$

The general form of the quadratic operator is

$$Q[x(n)] = \sum_m \sum_r c_{mr}\ x(n+m)x(n+r) \tag{9}$$

where c_{mr} are constants. For $Q = \Psi_d$ the time-frequency distribution obtained in Fig. 1 is the energy spectrum: $TF(n,k) = ES(n,k)$. For $Q(x) = x^2$ the time-frequency distribution obtained is the short-time smooth power spectral envelope[1] $TF(n,k) = PS(n,k)$ where

$$PS(n,k) = \sum_{m=n}^{n+N-1} [x_k(m)]^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_k(\omega)X(\omega)|^2 d\omega \tag{10}$$

assuming $x(n)$ is zero outside $[n, n+N-1]$.

Using Eqs. (7), (10) the ratio between the power spectral envelope and the energy spectrum can be approximated by

$$\frac{ES(n,k)}{PS(n,k)} \approx 1 - \cos(2\omega_k) = 2\sin^2(\omega_k). \tag{11}$$

The approximation is valid for narrowband signals $x_k(n)$, where the spectral energy $|X_k(\omega)|$ is concentrated around ω_k and the slowly-varying (in frequency) $\cos(2\omega)$ term can be assumed constant within the bandwidth of $x_k(n)$. Second-order approximations of Eq. (7), i.e., $ES(n,k) = (1/\pi) \int_{-\pi}^{\pi} \omega^2 |H_k(\omega)X(\omega)|^2 d\omega$, can be shown to cause formant spectral peak translation in addition to the scaling apparent in Eq. (11). Specifically, formant peaks with center frequencies up to $F_s/4$ Hz are translated towards the lower frequencies in the energy spectrum, and vice-versa for formant frequencies higher than $F_s/4$ (thus formant translation is a function of the sampling frequency F_s).

In Fig. 2, a time-slice of the ratio $\{ES(n,k)/PS(n,k) :\ k = 1..K\}$ is shown (solid line) together with the function $1 - \cos(2\omega_k)$ (dashed line). The ratio is computed for a single 20 ms speech frame of the vowel /ih/. A uniformly-spaced Gabor filterbank with 250-Hz 3-dB bandwidth per Gabor filter was used for computing ES and PS (sampling frequency 16 kHz). Differences between the computed and predicted ratio values are due to second-order effects (ripples in Fig. 2 correspond to formant translations in ES) and to the use of the (approximate) discrete Fourier transform instead of the discrete-time Fourier transform. Most ASR front-ends use the inverse cosine transform of the logarithm of $PS(n,k)$ as a feature set (cepstrum). In the cepstrum domain, the difference between energy cepstrum and "standard" cepstrum is approximately a time-independent bias.

In general, using Eq. (5) the sum of any quadratic operator $Q[x(n)]$ output (e.g., see [4, 1]) can be expressed as

$$\sum_{n=-\infty}^{\infty} Q[x(n)] = \sum_{\ell=0}^{L} \frac{c_\ell}{\pi} \int_{-\pi}^{\pi} \cos(\ell\omega)\ |X(\omega)|^2 d\omega \tag{12}$$

[1] For computational efficiency the spectral envelope $PS(n,k)$ is computed as $S(n,k) = (1/\pi) \int_0^{\pi} |X_k(\omega)|^2 d\omega$ rather than in the time domain as in Fig. 1.

where c_ℓ are arbitrary constants. For narrowband signals $x_k(n)$, $\cos(\ell\omega)$ can be assumed constant around ω_k and the short-time average of $Q[x_k(n]$ can be expressed as

$$log[TF(n,k)] - log[PS(n,k)] \approx b(k) \tag{13}$$

i.e., the difference between the log of any time-frequency distribution produced by the generalized ASR front-end in Fig. 1 and the log of the power spectral envelope is approximately a time-independent bias vector b (also in the cepstrum domain).

Given the similarity between the time-frequency distributions of quadratic operators it is expected that ASR performance will also be similar for various front-ends that use short-time averages of quadratic operators as features. However, as the size of the short-time window N decreases and/or the bandwidth of the filter $H_k(\omega)$ increases the differences among $TF(n,k)$ are no longer time-invariant, i.e., $b = b(n,k)$ and significant ASR performance differences may arise between various front-ends (see for example [12] where the energy operator is applied to the unfiltered signal). The equivalence between $ES(n,k)$, $PS(n,k)$ and $S(n,k) = (1/\pi)\int_0^\pi |X_k(\omega)|^2 d\omega$ as features (in the cepstrum domain) for ASR is experimentally shown in Section 4.

3 Spectral Moments and Average Instantaneous Frequency

In this section, we investigate the relation between various time-frequency distributions motivated by the AM-FM modulation model [3], the sinusoidal speech model [5], and spectral analysis. The distributions compute the short-time instantaneous frequency in different frequency bands. The distributions are compared to the short-time spectral envelope and their application to ASR is discussed.

The AM-FM modulation model, introduced in [3], describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure

$$r(t) = a^I(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \tag{14}$$

where $f_c \triangleq F$ is the "center value" of the formant frequency, $q(t)$ is the frequency modulating signal, and $a^I(t)$ is the time-varying amplitude. The instantaneous formant frequency signal is defined as $f^I(t) = f_c + q(t)$. The speech signal $x(t)$ is modeled as the sum $x(t) = \sum_{k=1}^K r_k(t)$ of K such AM–FM signals, one for each formant. A general family of time-frequency distributions of amplitude weighted short-time averages of the instantaneous frequency is defined as

$$TF_1(n,k) = \frac{\sum_{m=n}^{n+N-1}[a_k^I(n)]^\gamma f^I(n)}{\sum_{m=n}^{n+N-1}[a_k^I(n)]^\gamma} \tag{15}$$

where $a_k^I(n)$, $f_k^I(n)$ are the amplitude envelope and the instantaneous frequency, respectively, of the narrow-band signal $x_k(n)$ in Eq. (3), and γ is an arbitrary constant. Note that TF_1 for $\gamma = 0$, $F_u =< f^I(t) >$, was used for fundamental frequency estimation in [10] and TF_1 for $\gamma = 2$, $F_w =< [a^I(t)]^2 [f^I(t)] > / < [a^I(t)]^2 >$, (also referred to as the "pyknogram") was used for formant tracking in [9].

The sinusoidal model [5] models the speech signal $x(t)$ as a superposition of short-time varying sinusoids. Similarly the narrow-band signals $x_k(t)$ can be modeled using a sinusoidal model as

$$x_k(t) = \sum_{p=1}^{P_k} a_{kp} \; sin(2\pi f_{kp} t + \theta_{kp}) \tag{16}$$

where a_{kp}, f_{kp}, θ_{kp} are the constant (in an analysis frame $[n, n + N - 1]$) amplitudes, frequencies and phases, respectively, of the $p = 1 ... P_k$ sinusoids modeling $x_k(t)$. A general time-frequency representation can be obtained as a weighted average of f_{kp} as follows

$$TF_2(n,k) = \frac{\sum_{p=1}^{P_k} [a_{kp}(n)]^\gamma \; f_{kp}(n)}{\sum_{p=n}^{P_k} [a_{kp}(n)]^\gamma}. \tag{17}$$

where γ is an arbitrary constant. Note that the summation index p is a frequency index.

Finally a third type of time-frequency distribution is the generalized first spectral moment

$$TF_3(n,k) = \frac{1}{2\pi} \frac{\int_0^\pi |X_k(\omega)|^\gamma \; \omega \; d\omega}{\int_0^\pi |X_k(\omega)|^\gamma d\omega} \tag{18}$$

where γ is an arbitrary constant. Note that TF_3 for $\gamma = 0.5$ has been used as an ASR feature in [6]. Next we investigate the relationships among the three time-frequency distributions TF_1, TF_2 and TF_3 defined above.

Clearly TF_2 is a short-time estimate of the generalized spectral moment, i.e., $TF_3 \approx TF_2$. As P_k goes to infinity in Eq. (16) (i.e., more sinusoidal components are included in the approximation) the time-frequency representations TF_2, TF_3 become equal. The relation between TF_1 and TF_2 is more complicated and depends on the value of the amplitude weight γ. Specifically, for $\gamma = 2$, it is easy to show that all three time-frequency distributions are equivalent, i.e., $TF_1 = TF_2 = TF_3$ [9]. For $\gamma \neq 2$, one can show (along the lines of the proof for $\gamma = 0$ in [10]) that under the assumption that f_{kp} are harmonically related

$$TF_1(n,k) \approx \frac{[a_{kM}(n)]^2 f_{kM}(n) + \frac{\gamma}{2} \sum_{p \neq M} [a_{kp}(n)]^2 f_{kp}(n)}{[a_{kM}(n)]^2 + \frac{\gamma}{2} \sum_{p \neq M} [a_{kp}(n)]^2} \tag{19}$$

where $a_{kM} = \max_p \{a_{kp}\}$ is the amplitude of the sinusoid f_{kM} with the greatest amplitude. Thus, *we have established that* TF_1, TF_2, TF_3 *are equivalent for* γ *around 2.* Next we investigate the relationship between TF_3 and the standard ASR front-end.

The standard ASR front-end computes the short-time spectral energy in each of the frequency bins k as follows: $S(n,k) = (1/\pi)\int_0^{\pi} |X_k(\omega)|^2 d\omega$, where $X_k(\omega)$ is defined in Eq.(3). Assuming that $h_k(n)$ in Eq.(3) is the real Gabor filter's impulse response, the frequency response can be expressed as

$$H_k(\omega) = (\sqrt{\pi}/2\alpha)(e^{-(\omega-\omega_k)^2/4\alpha^2} + e^{-(\omega+\omega_k)^2/4\alpha^2}) \tag{20}$$

where α is proportional to the bandwidth of the filter. For $\gamma = 2$ and for a Gabor filterbank, the spectral moment time-frequency distribution TF_3 can be expressed as a function of the standard front-end feature set S as follows[2]

$$TF_3(n,k) \approx \omega_k + \frac{\alpha^2}{S(n,k)} \frac{dS(n,k)}{d\omega_k}. \tag{21}$$

where $dS(n,k)/d\omega_k$ is the derivative of the short-time spectral energy distribution S with respect to the center frequency of the filterbank filter ω_k.

Given the close relationship between TF_3 and S it might be expected that both distributions will perform similarly when used as features for ASR. However, S is a zeroth-order spectral estimator while TF_3 is a first-order one (see Eq. (18)). Thus, TF_3 is expected to be a less robust estimator and have inferior classification performance. Indeed, we have experimentally verified that the separability of phonemic classes in the $S(k)$ space is significantly better than in the $TF_3(k)$ space. Efforts to augment the standard feature set S by one of TF_{1-3} are expected to have little success [6] due to the high correlation between the two feature sets exemplified by Eq. (21). Note, however, that gains may be observed when different analysis time-scales are used for the two distributions or for mismatched ASR conditions (in training and testing), e.g., noisy speech. Further, since $TF_1 \approx TF_3 \approx TF_2$ for $\gamma \approx 2$ the above statements are also valid for TF_1 and TF_2.

4 Experiments

In this section, the recognition accuracy of the various feature sets is compared for a connected digit recognition task[3]. A hidden Markov model (HMM) recognizer was used with 8 Gaussian mixtures per HMM state. Each digit was modeled by a left to right HMM unit, 8-10 states in length. The test set consists of 4304 digit strings (13185 digits) collected over the public switched telephone network.

The front-ends evaluated were (all with 20 ms analysis window, 10 ms update, and identical filterbank spacing and bandwidths): (1) "standard" mel-filterbank front-end $S(n,k)$ using triangular

[2] The approximation error is greatest for ω_k close to 0 and for large values of bandwidth parameter α.

[3] Similar results were obtained on the TIMIT phone recognition task.

filters, (2) mel-filterbank front-end using Gaussian filters $PS(n,k)$, (3) energy spectrum $ES(n,k)$, and (4) amplitude weighted average instantaneous frequency $TF_1(n,k)$ for $\gamma = 2$. For all front ends the feature set consisted of the mean square of the signal ("standard energy"), the inverse cosine of the above described time-frequency distributions (cepstrum), and the first and second derivatives of these features.

The results are shown in Table 1. As expected the performance of S, PS and ES is very similar, while TF_1 performs significantly worse. This is consistent with the theoretical results obtained in Sections 2 and 3.

5 Conclusions

We have established the close relationship among various short-time distributions and provided baseline results comparing the ASR performance of these alternative feature sets with the standard ASR front-end. Specifically, it was shown that: (i) the difference between cepstum ASR features derived from short-time averages of quadratic operators and the standard ASR front-end is a time-independent bias, provided that identical time-frequency tiling and narrowband filters are used in the ASR front-end, (ii) TF_1, TF_2 and TF_3 are equivalent time-frequency representations when amplitude squared weighting is used ($\gamma = 2$), and can be expressed as the derivative of the spectral energy distribution S. The implications of these results for speech recognition were also discussed and experimentally verified. For matched training and testing conditions, ASR front-ends using cepstrum derived from averages of quadratic operators were shown to perform similarly to the standard ASR front end, while front-ends using first spectral moment features were shown to perform significantly worse.

References

[1] L. Atlas and J. Fang, "Quadratic detectors for general nonlinear analysis of speech," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, (San Francisco, CA), pp. II: 9–12, Mar. 1992.

[2] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, (Albuquerque, New Mexico), pp. 381–384, Apr. 1990.

[3] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[4] P. Maragos and A. Potamianos, "Higher–order differential energy operators," *IEEE Signal Processing Letters*, vol. 2, Aug. 1995.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[6] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, (Seattle,Washington), pp. 617–620, May 1998.

[7] J. W. Pitton, K. Wang, and B. H. Juang., "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proceedings of the IEEE*, vol. 84, pp. 1199–1214, Sept. 1996.

[8] A. Potamianos and P. Maragos, "Applications of speech processing using an AM–FM modulation model and energy operators," in *Proc. European Signal Process. Conf.*, (Edinburgh, Scotland), pp. III: 1669–1672, Sept. 1994.

[9] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of the Acoustical Society of America*, vol. 99, pp. 3795–3806, June 1996.

[10] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech Communication*, vol. 28, pp. 195–209, 1999.

[11] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 599–601, Oct. 1980.

[12] H. Tolba and D. O'Shaughnessy, "Automatic speech recognition based on cepstral coefficients and a mel-based discrete energy operator," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process.*, (Seattle,Washington), pp. 973–976, May 1998.

[13] G. Zhou, J. Hansen, and J. F. Kaiser, "Linear and nonlinear speech feature analysis for stress classification," in *Internat. Conf. Speech Language Processing*, (Sydney, Australia), pp. 840–843, Dec. 1998.

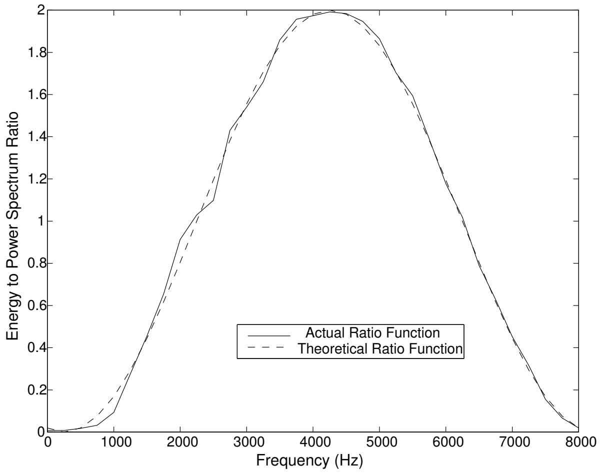

Figure 1: Ratio of energy spectrum over power spectral envelope.

Front End	Fltbnk	W.Acc.
Baseline cep: $C^{-1}\{S\} = c_1..c_{12}$	triang.	97.1 %
Power cep: $C^{-1}\{PS\}$	Gabor	96.8 %
Energy cep: $C^{-1}\{ES\}$	Gabor	97.0 %
Pyknogram cep: $C^{-1}\{TF_1\}$	Gabor	81.1 %

Table 1: Digit error rate for different time-frequency distributions as ASR feature sets (C^{-1} is the inverse cosine transform).