CROSS-SPECTRAL BASED FORMANT ESTIMATION AND ALIGNMENT

D.J. Nelson

U.S. Dept. of Defense
Ft. Meade, MD 20755-6514

ABSTRACT

We develop new algorithms, which address the problems of accurate frequency estimation and the comparison of the frequency content of two different spectra. We apply these algorithms to speech to accurately estimate formants from very little data and to estimate the frequency differences of individual formants of similar vowel instances. The first algorithm a simple method for construction of a very high-resolution, high-gain spectrum. This phase-re-parameterized spectrum is based on a deterministic signal model and uses both Fourier phase and magnitude and accounts for all of the energy in the Fourier spectrum. To compare the formants of similar vowels we introduce a spectral correlation algorithm based on an outer product rather than the inner product of conventional correlation.

1. INTRODUCTION

Since the classical paper by Peterson and Barney [1], there have been many studies of the formant structures of phonetically similar sounds spoken by different speakers. Examples of these studies are the validation of the Peterson and Barney experiment by Hillenbrand, et al [2] and the estimation of the speech scale of Umesh et al, in which they estimate a spectral warping function, under which the spectra of similar vowels spoken by different speakers differ by essentially a translation [3, 4, 5]. All of these processes, to date, have required the use of skilled humans to estimate the formant frequencies and align the spectra for comparison.

We develop methods which may be used to accurately estimate formant frequencies of individual vowels and to accurately estimate the frequency differences of formants of similar vowels spoken by different speakers. These methods are simple, accurate and may be easily automated. We base these methods on previously reported results (c.f. [6]). The basis of these methods is a phase differentiation of the Fourier spectrum, which was used in the early 1980's in a variety of signal processing algorithms and was published as the cross-power-spectrum (CPS) [7, 8]. In adapting these methods to speech formant estimation, a deterministic resonance model is used and a cross-spectral representation is computed from very short frames of data (typically 3 - 5 ms). The ability to accurately estimate spectral components from such short frames makes these methods ideal for processing non-stationary signals, such as speech. A comprehensive explanation of the cross-spectral process and the deterministic resonance model is presented in a paper published in JASA [6].

To construct an accurate and very high gain spectral representation, we apply the re-mapping proposed by Nelson (c.f. [6]) and then uniformly re-sample the resulting spectrum. We call the resulting spectral representation a phase-re-parameterized (PR) spectrum. This process is very efficient and accounts for all of the energy in the Fourier spectrum. While we demonstrate this method on the estimation of the formants of near steady state vowels, these methods have been successfully applied to the general problems of formant estimation and tracking, the estimation of the excitation fundamental F0 of speech and a variety of speech and communication problems.

As an application of the PR spectrum, we address the problem of estimating the frequency differences of the formants of two similar vowels. In this application, a correlation function based on an outer product is proposed. In this process, an outer product of the PR spectra of the two vowels is computed and the frequency differences of the formants are computed from the argument or angle of the resulting product. This provides an accurate comparison of individual formants, in contrast to conventional correlation methods which can only provide an ensemble frequency difference estimate.

In the implementation presented here, we process 3 ms Hanning windowed frames of vowels from Hillenbrandt's Western Michigan data consisting of 139 speakers, each saying 12 vowels in an hVd context [9]. This short frame length was used to insure that there were no pitch artifacts [6]. In addition, a pitch-synchronization process was used to select frames centered approximately 5 ms after glottal closure. This synchronization process is fully automated and is based on the phase-based resonance and excitation indicators proposed by Nelson [6].

2. THE SIGNAL MODEL AND THE CROSS-SPECTRAL REPRESENTATION

We base our spectral estimation and correlation methods on a signal model as the sum of a sparse set of AM/FM modulated analytic components

$$f(t) = \sum_n f_n(t) = \sum_n a_n(t) e^{i\phi_n(t)}, \qquad (1)$$

where the instantaneous frequency of the n^{th} component is $\omega_n(t) = (d/dt)\phi_n(t)$.[1] The model represented by Eq. (1)

[1] We stress that that instantaneous frequency of signal components is not the same as frequency in the Fourier sense. It is based on a model in which the analytic signal is distributed. This

is a deterministic signal model, in which each signal component has a well defined amplitude and frequency at each time. We further assume that the frequencies, $\omega_n(t)$, of the signal components change very slowly with time, while the AM components, $a_n(t)$, are more rapidly changing. This is consistent with a speech resonance model in which the formants (vocal tract resonances) change as the position of the tongue and configuration of the vocal tract change, while the AM modulation reflects changes in energy during a glottal excitation cycle. While there is a significant change in energy during the glottal excitation cycle, the configuration of the vocal tract may be assumed to change very little over one cycle. If the analysis interval is less than the period of an excitation cycle, there can be no observed excitation harmonics. By assuming that instantaneous frequencies are slowly varying, we reduce the problem from time-frequency analysis to an analysis of frequency alone. Furthermore, the methods we present here generalize to non-stationary components. For a full discussion of this and the absence of excitation harmonics, c.f.[6].

We start with the short time Fourier transform (STFT) [10]

$$\mathbf{F}(\omega, T) = \int_{-\infty}^{\infty} f(t+T)w(t)e^{-i\omega t}dt, \qquad (2)$$

where $w(t)$ is an analysis window, assumed to be short. For the examples presented here, a Hanning window of length 3 to 5 milliseconds was used. The dependence of $\mathbf{F}(\omega, T)$ on $w(t)$ is assumed, but is omitted from the notation for simplicity. A cross-spectrum is then computed as the product of the STFT and the complex conjugate of the STFT of the signal delayed in time (typically by one sample). The argument of the cross-spectrum is essentially the derivative, with respect to time, of the phase of the STFT.

The channelized instantaneous frequency (CIF) is defined as

$$\mathbf{CIF}_f(\omega, T) = \frac{\partial}{\partial T} \arg\{\mathbf{F}(\omega, T)\} \qquad (3)$$

It is convenient to encode the CIF as the phase of a complex valued surface

$$\mathbf{C}_f(\omega, T) = A_f(\omega, T)e^{i\mathbf{CIF}_f(\omega,T)} \qquad (4)$$

where it is understood that these functions depend on $A(\omega, T)$

We assume $0 \leq \omega \leq \pi$ and $A_f(\omega, T)$ is a real function, which is normally related to the magnitude of $\mathbf{F}(\omega, T)$. For the speech applications presented here, we have used

$$A_f(\omega, T) = \max(0, 20\log_{10}\frac{|\mathbf{F}(\omega, T)|}{M} + \tau), \qquad (5)$$

where M is the global absolute maximum of $\mathbf{F}(\omega, T)$ for the speech utterance, and τ is a threshold, dependent on SNR, which is normally set at 50 dB.

Note that there are two frequency representations encoded in Eq. (4). The observed frequency is ω. For an observation made at (ω_0, T_0), the estimated instantaneous signal frequency is $\arg\{\mathbf{C}_f(\omega_0, T_0)\}$ (c.f.[6]).

The cross-spectral representation, is based on a distribution of the analytic signal, and not energy, in time and frequency. For deterministic narrowband FM components, as we have modeled them, the frequency of each component is a well defined function of time. If the signal components are isolated in time-frequency, each observed spectral component, $\mathbf{F}(\omega_0, T_0)$, represents an observation of one signal component, as it exists at one instant in time. The actual time and frequency instants of the signal component may not be the same as the time and frequency at which the observation is made. By assuming that the actual location of the signal component is consistent with the CIF we may significantly improve the accuracy of the estimated signal frequency (c.f.[6]). This principle is the basis of spectral estimation techniques, such as the instantaneous frequency [11], Kay's method [12], the cross-power spectrum [7, 8].

concept is not the same as the conventional notion of frequency as an energy distribution.

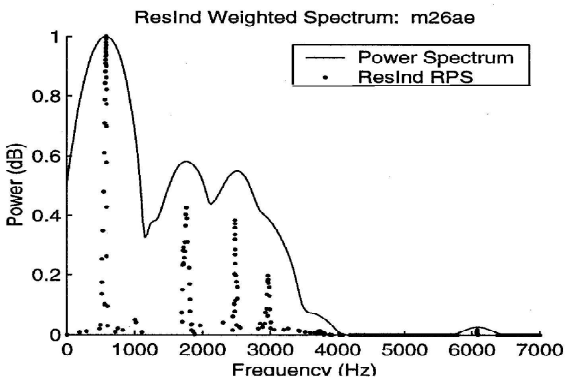

Fig. 1. Pitch synchronous spectra. Log power spectrum (solid line) and resonance indicator weighted log power spectrum with CIF as abscissa. Synchronization based on resonance indicator [6].

It should be noted that, in the examples presented here, we have employed a pitch synchronization process. Each spectrum displayed was computed from a single 3 ms to 5 ms frame centered approximately 4 ms after glottal closure. The synchronization process is fully automated and is based on the phase-based excitation and resonance indicator surfaces proposed by Nelson [6]. To estimate time of glottal closure and maximal resonance, the indicator surfaces are simply averaged with respect to frequency.

3. RE-PARAMETERIZATION BY PHASE

In representation Eq. (4), the index frequency, ω, and the instantaneous frequency representations are not in agreement with each other, in the sense that they indicate different frequencies. We would like to compute a single spectrum in which both frequency representations are in agreement. Since the signal frequency is assumed to be best approximated by the CIF, it is that representation we wish to preserve.

As noted, $\mathbf{C}_f(\omega, T)$ is assumed to have the angular instantaneous frequency estimates of a narrowband signal component encoded in its argument. In Eq. (4), ω represents a parameterization of the angular Fourier frequency, and the argument, $\mathbf{CIF}_f(\omega, T)$, represents the instanta-

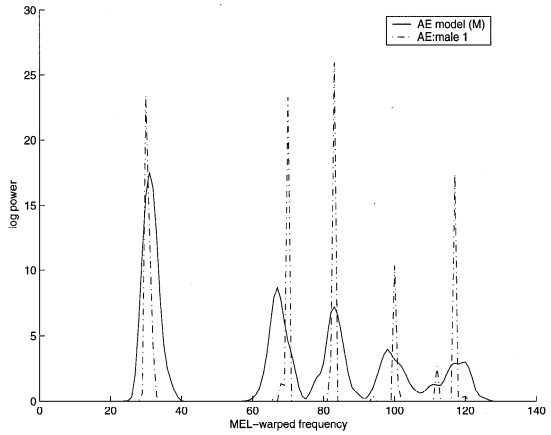

Fig. 2. A single male vowel AE, re-parameterized by phase and the average of 45 similar male vowels re-parameterized by phase

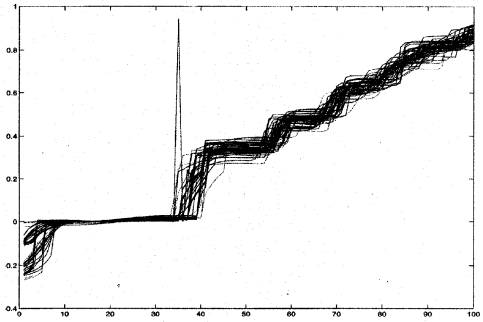

Fig. 3. The CIF for vowels AE, spoken by 45 adult male speakers. The nominal frequency of each first formant has been translated to zero for illustration. Each step represents one formant.

neous angular signal frequency. $\mathbf{CIF}_f(\omega, T_0)$ is a nonlinear, non-monotonic function of ω, which may be expected to cluster near frequencies $\{\omega_n(T_0)\}$ of the narrowband signal components at time T_0 (where we have ignored the effects of group delay for simplicity). We would like to re-parameterize the frequency axis of the surface so that the representation is linear in instantaneous frequency. Inversion is not possible, since $\mathbf{CIF}_f(\omega, T_0)$ is not assumed to be monotonic in ω. We can, however, accumulate the surface values which assume each value of $\mathbf{CIF}_f(\omega, T_0)$.

$$\mathbf{P}_f(\Phi, T_0) = \sum_{\omega \in \omega_\Phi} \frac{\mathbf{C}_f(\omega, T_0)}{1 + |d\mathbf{CIF}_f(\omega, T_0)/d\omega|}, \qquad (6)$$

where $\omega_\Phi = \{\omega | \mathbf{CIF}_f(\omega, T_0) = \Phi\}$. For values of $\mathbf{CIF}_f(\omega, T_0)$, which are not observed, we assign a phase value equal to ω and a very small amplitude value approximately equal to the minimum positive number supported by the machine precision. The factor in the denominator of Eq. (6) is a correction needed to preserve energy due to the spectral warping in the representation. In Eq. (6) we have used the summation notation since the we expect ω_Φ to be sparse. A typical example of re-parameterization by phase is depicted in Fig. 2. The basis for the processing gain for formant estimation is the experimentally derived observation that $\mathbf{CIF}_f(\omega, T_0)$ is nearly constant for all values of ω within the same formant band. This observation is depicted in Fig. 3, The fact that the CIF is nearly constant across each formant band is an indication that the deterministic signal model is valid for speech.

The concept of re-parameterizing by phase is not a convolution or smoothing in the normal sense. In smoothing, the value of the output is computed as the weighted average of the function evaluated in a neighborhood of a point. In re-parameterizing, we do not consider neighboring points. We merely ask where the point re-maps under the CIF, and we add its contribution to that new location. This phase-re-parameterized (PR) spectral representation produces extremely narrow, high-gain formant representations from as little as 3 ms of data.

4. OUTER PRODUCT AND CORRELATION

Our purpose in developing the PR spectrum was to accurately estimate formant frequencies. We now turn our attention to the problem of estimating the frequency differences of individual formants of similar vowels spoken by different speakers. To accomplish this, we introduce a correlation function based on an outer product, in contrast to the normal cross-correlation function, which is computed as the dot or inner product of the two spectra at a sequence of frequency lags. The normal cross-correlation function may produce an estimate of the frequency lag for which the two spectra are best aligned, but it is an ensemble estimate, which provides no information about the lags which best align the individual spectral (formant) components.

It may be argued that we could simply estimate the n^{th} formants of the two vowels and compute their frequency difference by subtracting their frequencies. This would work, if there are no missing formants and no spurious spectral components. In this case, the formants may not be correctly identified. We therefore propose the following correlation method.

For two PR spectral surfaces, $\mathbf{P}_{f_0}(\omega, T_0)$ and $\mathbf{P}_{f_1}(\zeta, T_0)$, we may define their outer product at time T_0 as the surface

$$\mathbf{OP}_{f_0 f_1}(\omega, \zeta, T_0) = \mathbf{P}_{f_0}(\omega, T_0)\mathbf{P}^*_{f_1}(\zeta, T_0). \qquad (7)$$

The outer product represents the entire correlation properties of the spectra $F_0(\omega, T_0)$ and $F_1(\zeta, T_0)$. The conventional frequency-lag correlation function may be obtained by integrating $\mathbf{OP}_{f_0 f_1}(\omega, \zeta, T_0)$ along diagonals

$$\mathbf{R}_{f_0 f_1}(\tau, T_0) = \int_{\omega - \zeta = \tau} \mathbf{OP}_{f_0, f_1}(\omega, \zeta, T_0) \qquad (8)$$

If the magnitude of the outer product Eq. (7) is large for some ω_0 and ζ_0, the magnitudes of both $\mathbf{P}_{f_0}(\omega_0, T_0)$ and $\mathbf{P}_{f_1}(\zeta_0, T_0)$ must be large. In this case, we may conclude that $\arg\{\mathbf{OP}_{f_0 f_1}(\omega_0, \zeta_0, T_0)\}$ represents a valid frequency

difference of signal components of f_0 and f_1, observed at (ω_0, T_0) and (ζ_0, T_0) respectively. If the PR spectra are impulsive (sparse), $\mathbf{OP}_{f_0 f_1}(\omega, \zeta, T_0)$ is impulsive.

Finally, we may compute a single frequency-lag correlation function by re-parameterizing the outer product, $\mathbf{OP}_{f_0 f_1}(\omega, \zeta, T_0)$ by phase. Under this operation, outer product components with similar phase are added, as in Eq. (6) to produce a single frequency-lag outer-product based correlation function. The resulting complex-valued correlation function is convolved with at smoothing windowed. Since each complex correlation component retains knowledge of the frequency difference it represents, the resulting smoothed PR correlation function effectively averages the contributions of formant pairs, whose offsets are approximately the same. This smoothed correlation function, depicted in the bottom trace of Fig. 4, has similar properties to the conventional correlation function, but it is phase based and impulsive, producing much more accurate frequency lag estimates and much higher gain than the conventional correlation function.

5. CONCLUSIONS

We have presented new spectral analysis and correlation techniques which provide greatly improved gain and accuracy over conventional methods. These algorithms have been fully automated and applied to the problem of resolving and estimating speech formants. In addition, these methods provide the capability to easily estimate accurate frequency differences of corresponding formants of similar vowels spoken by different speakers.

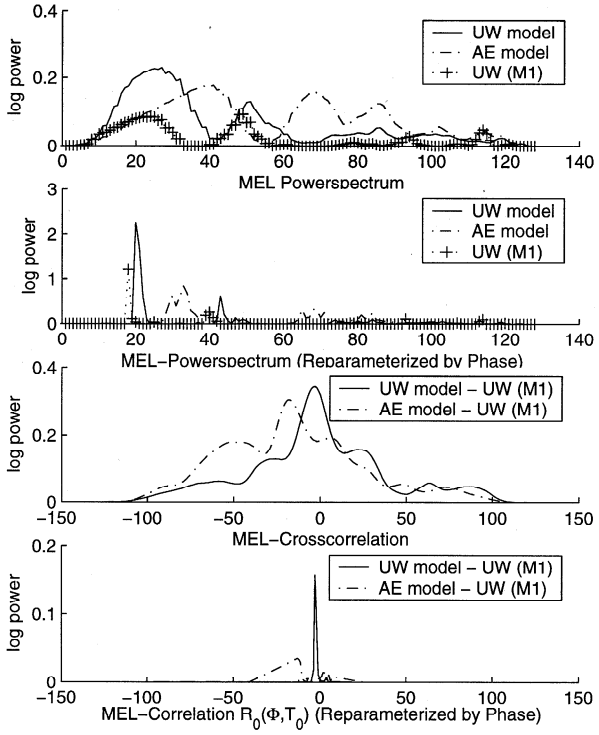

Fig. 4. Conventional cross-correlation and outer product re-parameterized by phase, represented on a MEL scale. Upper trace: A single male vowel "UW" and two vowel models computed by averaging aligned vowels. Second trace: PR power spectra. Third trace: conventional cross-correlation of single "UW" and two vowel models. Bottom trace: Outer product, re-parameterized by phase.

6. REFERENCES

[1] G. E. Peterson and H.L. Barney, "Control methods used in a study of the vowels," in Journal of the Acoust. Society of America, vol. 24, pp. 174-184, 1952.

[2] James Hillenbrand et al.,"Acoustic characteristics of American English vowels," in Journal of the Acoust. Society of America, vol. 97 no. 5, pp. 3099-3111, May 1995.

[3] S. Umesh, L. Cohen, D. Nelson, "Improved Scale-Cepstral Analysis in Speech," Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing,Seattle, 1998.

[4] S. Umesh, L. Cohen, D. Nelson, "Frequency Warping and The MEL Scale," IEEE Signal Processing Letters, January 2001.

[5] S. Umesh, L. Cohen, D. Nelson, "The speech scale," Acoust. Soc. Amer. Elect. Letters, Accepted Jan. 10, 2002.

[6] D.J. Nelson, "Cross-Spectral Methods for Processing Speech," in Journal of the Acoustic Society of America, November 2001.

[7] D.J. Nelson,"Special Purpose Correlation Functions," internal technical report, 1986.

[8] D.J. Nelson,"Special Purpose Correlation Functions for Improved Parameter estimation and Signal Detection," Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, Minneapolis, pp. 73-76, April,1993.

[9] J. Hillenbrand, Western Michigan Vowel Data, 1995.

[10] D. Gabor, "Theory of Communication," Proc. of the Inst of Elect Eng., vol. 93, no.26, pp 429-457, 1946.

[11] A.J. Gibbs, "The Design of Digital Filters," in Australian Telecommunication Research Journal, vol 4, pp 29-34, 1970, reprinted in Digital Signal Processing, ed L.R. Rabiner and C.M. Rader IEEE Press, pp.35-42, 1972.

[12] S. M. Kay, "Statistically/Computationally Efficient Frequency Estimation," Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, pp. 2292-2295, 1988.