Correlation Based Speech Formant Recovery

Douglas Nelson
Department of Defense
Fort George G. Meade, Md. 20755
djnelso@afterlife.ncsc.mil

ABSTRACT

A new method for generating speech spectrograms is presented. This algorithm is based on an autocorrelation function whose parameters are chosen provide processing gain and formant resolution, while minimizing pitch artifacts in the spectrum. Crisp formants are produced, and the power ratio of the formants can be adjusted by pre-filtering the data. The autocorrelation process is functionally equivalent to a time-smoothed, windowed Wigner distribution. The process is an improvement over the normal FFT implementation since it requires much less data to resolve the speech formants, and it is an improvement over the un-smoothed Wigner distribution since the cross-terms normally associated with the Wigner distribution are greatly attenuated by the smoothing operation.

1. INTRODUCTION

In machine processing of speech, the front end is a process which produces time varying feature vectors from the speech data. These feature vectors are assumed to represent the speech well enough that the desired information may be extracted statistically from them. The most common speech front end is the Mel-warped cepstrum. In computing the Mel-warped cepstrum, a powerspectrum is first computed, using a narrowband FFT based on a large analysis window. The powerspectrum is then smoothed and resampled to approximate a Mel-warped spectrum which is approximately linearly sampled below 1 kHz and exponentially sampled (log warped) above 1 kHz. The cepstrum is then computed as

$$C(\zeta) = \int \log\left|\hat{S}(M(\omega))\right| \cos(2\pi\zeta\omega)\, d\omega, \qquad 1$$

where $\hat{S}$ is the Fourier spectrum of the signal and M is the Mel warping function. To complete the process, the cepstrum is truncated.

For a speech signal sampled at 8 kHz, the initial spectrum is typically computed using a 256 point FFT, and the cepstrum is truncated to approximately 12 coefficients. A transform of this size provides enough resolution to resolve speech formants fairly well, but it oversamples the speech spectrum since the pitch harmonics are also resolved. In machine processing of speech, sensitivity to pitch is generally an undesirable property of front-ends, since it results in poor correlation of the information-bearing formants. The purpose of spectral smoothing and the cepstral truncation is to mitigate the effects of pitch. The action of the cepstral truncation is subtle since it is functionally equivalent to smoothing the Mel-warped spectrum by performing a circular convolution of the Mel-warped power spectrum with a sinc function.

An alternative to smoothing the narrowband power spectrum is to base the process on wideband FFT, using a small analysis window. This process has the advantages that it provides fine time resolution and that pitch is not apparent in the wideband spectrum, but it has the disadvantage that the short analysis interval affords much less processing gain and poor formant resolution.

Time-frequency distributions, such as the Wigner distribution could be used to produce a representation with most of the desired properties, but such quadratic processes have the distinct disadvantage that multi-component signals, such as speech produce Wigner distributions with undesirable cross-terms.

The methods of this paper produce a spectral representation of speech which contains the formants but almost no artifacts of the excitation functions. The process has high gain associated with the narrowband spectrogram, but the formant structure has much finer definition. The smoothing operation in this process is performed by a linear autocorrelation function which is computed from the raw time domain signal. As was observed by Leon Cohen, this autocorrelation-based process is functionally equivalent to a time-smoothed Wigner distribution. However, unlike the normal time-frequency distributions, no cross-terms are produced.

2. STRUCTURE OF SPEECH

Speech is the result of the excitation of the vocal tract by a quasi-periodic glottal pulse train and/or frication caused by a constriction of the air flow in the front cavity of the vocal tract. The signal components which are excited by glottal pulses are called voiced or pitch related, since pitch is the frequency of the glottal pulses. The speech signal is functionally equivalent to a convolution of the impulse response of the vocal tract and the excitation functions.

$$S(t) = V_p(t) * P(t) + V_F(t) * F(t) , \qquad (2)$$

where $S(t)$, $P(t)$ and $F(t)$ are respectively the received signal, the pitch excitation function and the frication excitation function, and $V_p(t)$ and $V_F(t)$ are the impulse responses of the portions of the vocal tract excited by pitch and frication respectively.

The pitch excitation function is quasi-periodic with an expected fundamental frequency of approximately 50 to 250 Hz. Frication is a noise-like pseudo-random function, which excites primarily the speech spectrum above 1kHz. The formants are the spectral bulges representing the resonant structure of the vocal tract and are generally higher than 500 Hz.

For voiced speech, the time-domain representation of speech may be modeled as a relatively narrow glottal pulse followed by a series of reflections, as the vocal tract resonates at its natural frequencies. The primary reflection corresponds to the first formant and is generally shorter than 2 milliseconds (higher than 500 Hz.) In addition to the primary reflection of the glottal pulse, reflections are produced recursively as each reflection is in turn reflected in the vocal tract. The resulting structure is exponentially damped and locally multiply periodic, with the periodicity interrupted by the occurrence of the next glottal pulse. Similarly, the higher frequency formants appear as locally periodic structures in the time domain.

In the frequency domain, for voiced speech, the structure consists of the product of the pitch spectrum and the formant spectrum. The pitch spectrum consists of pitch bars which are spectral bulges at the harmonics of pitch. The formant spectrum consists of several fairly broad spectral bulges. The resulting spectral structure of voiced speech is a formant spectrum which is "sampled" at the pitch harmonics.

3. THE PROCESS

Removing pitch artifacts from the observed spectrum can be accomplished by designing a process in which the effective analysis interval contains no more than one glottal pulse. The observed spectrum is then essentially the impulse response of the vocal tract. Under the one pulse assumption, pitch can not be represented in the spectrum since this would require periodicity, and there can be no periodicity with only one pulse. There can, however, be non-causal formants which result when the reflections of an earlier glottal pulse interact with a later glottal pulse.

If we compute the windowed symmetric autocorrelation function

$$R_{tT}(\tau) = \int_{-T/2}^{T/2} w(\zeta) S\left(t+\zeta+\frac{\tau}{2}\right) S^*\left(t+\zeta-\frac{\tau}{2}\right) d\zeta \qquad 3$$

and restrict $|\tau|$ to be less than the expected pitch interval, then $R_{tT}(\tau)$ will be an even function with no pitch components, but will preserve the formant structure.

The autocorrelation function contains the same spectral information as the original signal, since the spectrum of the autocorrelation function is essentially the powerspectrum of the original signal. For voiced speech, the expected spectrum of $R_{tT}(\tau)$ is real and positive and even.

As a fine point, the spectrum of $R_{tT}(\tau)$ is not the power spectrum of the signal S since the convolution in (3) is linear, and the power spectrum is the spectrum of the circular autocorrelation function. Note that there is a processing gain of 3 dB/octave in the correlation function, and all of the spectral components of the formants we wish to recover are phased by the correlation function so that

$$E\left(\arg\left(\hat{R}_{tT}(\omega)\, e^{i\omega\tau}\right) \Big|_{\tau=0} \right) = 0 , \forall\omega \qquad (4)$$

If we compute the Fourier spectrum of (3), we get

$$\hat{R}_{tT}(\omega) = \int \int_{-T/2}^{T/2} S\left(t+\zeta+\frac{\tau}{2}\right) S^*\left(t+\zeta-\frac{\tau}{2}\right) d\zeta e^{i\omega\tau} dt , \qquad 5$$

which, after interchanging the order of integration, is seen to be the time smoothed Wigner distribution.

One of the functions of the autocorrelation function applied to speech is that it synchronizes the formant components. The expected formant contribution to the autocorrelation function is even, resulting in an expected zero phase Fourier spectrum. As a consequence, the autocorrelations, or equivalently their fourier transforms can be averaged as complex vectors to recover the processing gain of the narrow-band Fourier transform, without producing pitch components. The formant spectrum is then recovered as the half-wave rectified cosine transform of the averaged autocorrelation function

$$\hat{S}_{formant} = \frac{\hat{S} + |S|}{2} , \qquad 6$$

where

$$\hat{S} = \int_{\tau > 0} R_{tT}(\tau)\cos(2\pi\tau\omega)\,d\tau \quad . \qquad 7$$

4. CAUSAL .vs. NON-CAUSAL REFLECTIONS

In the previous section, we outlined a process for recovering the speech formant spectrum. The process is based on a short analysis window, and the desired spectral components are the result of an impulse exciting the vocal tract, which then resonates at its natural frequencies. As a word of warning, there is another situation which can produce spectral components. This is a non-causal situation in which the vocal tract's response to an earlier pulse interacts with a later pulse. This situation can cause spectral components which could be a problem if the analysis window is very short.

In the causal case a single glottal pulse excites a resonance. In the absence of other components, the phase of the resulting spectral component is determined by the glottal pulse. If the pulse occurs at time $t = t_0$ and the reflections occur at times

$$t_0 + m\Delta T,\ m = 1, 2\ldots, \qquad 8$$

then the spectrum will contain a bulge at

$$f = c/\Delta T \qquad 9$$

with phase

$$\varphi = \left.\frac{2\pi t_0}{\Delta T}\right|_{2\pi} . \qquad 10$$

Applying the autocorrelation function results in an expected phase whose value is zero.

To analyze the non-causal case, we consider an analysis interval spanning glottal pulse n and glottal pulse n + 1 , If we assume that the glottal pulses occur at time $t = t_n$, where n is the index of the glottal pulses, then the interaction of the reflections of pulse n with pulse n+1 itself produces a spectral component with the same frequency as the causal case, but the phase is different.

The non-causal bulges of the autocorrelation function occur at

$$t_{n+1} - (t_n + m\Delta T)\ ,\ m = 1, 2\ldots, \qquad 11$$

The frequency is the same as in the causal case, but the phase of the spectrum of the autocorrelation function is

$$\varphi_N = \left.\frac{2\pi(t_{n+1} - t_n)}{\Delta T}\right|_{2\pi} . \qquad 12$$

By computing the spectrum directly as the Fourier transform of an analysis interval consisting of several pulses, the spectral contributions destructively combine if the interval between pulses is not a multiple of the reflection period (i.e. the formant frequencies are not harmonics of pitch). Since pitch and formant frequencies are unrelated, the resulting spectral representation is suboptimal.

The autocorrelation method is still suboptimal, but the situation is better. Causal reflections always produce a spectral component with zero phase, and non causal reflections have random phase. By averaging, the non-causal contribution is mitigated. The contributions of the non-causal components can be further reduced by half-wave rectifying the cosine transform in equation (6). The rectified cosine-transformed spectrum has a processing gain due to the fact that some non-causal information is discarded with no loss of causal information.

5. A FINAL NOTE

The process outlined here produce good spectral representations of speech formant structures. The primary gain is that the process has good gain and resolves formants with less data than is necessary in normal powerspectrum methods. This is demonstrated in figures 5-10.

There are many methods for estimating the spectrum. It should be noted that the use of a Morlet wavelet frame instead of a Fourier transform has generally resulted in a spectrum with better definition in the higher formants. This was not developed in this paper since care must be taken in the construction of the basis. All figures in this paper were produced using the Fourier transform and not a wavelet transform.

REFERENCES

[1]Flanagan, J.L.,Speech Analysis Synthesis and Perception, Springer-Verlag, Berlin, 1965.

[2]Harris, J.D. and D.J. Nelson,"Glottal Pulse Alignment in Voiced Speech for Pitch Tetermination",Proc. IEEE-ICASSPConf., Vol.2,p519-22, 1993.

[3]Hartwick, J. Chang, D.Y., and Lim, J.S.,"Speech Enhancement Using the Dual Excitation Speech Model", Proc IEEE-ICASSP Conf., Vol.2,p367-370, 1993.

[4]Holton, T. and Love, S. "Robust Pitch and Voicing Detection using a Model of Auditory Signal Processing",Proc. Speech Research Symposium XIII at Hopkins University, June 1993.

[5]Nelson, D. "The Mellin-Wavelet Transform",Proc IEEE-ICASSP Conf., 1995.

[6]Pencak, J. and Nelson, D., "The NP Speech Activity Detection Algorithm",Proc IEEE-ICASSP Conf., 1995.

[7]Umesh, S., Cohen, L. Marinovic, N. and Nelson, D. "Scale transform in speech analysis",IEEE Trans on Speech and Audio Processing, 1996. Submitted

[8]Umesh, S., Cohen, L. Marinovic, N. and Nelson, D. "Frequency Warping and Speaker Normalization",. Proc IEEE-ICASSP Conf., 1996.Submitted

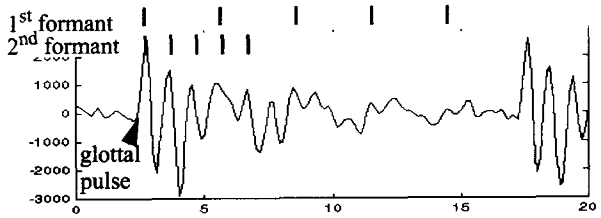

Figure 1 Typical voiced speech waveform, showing the periods of the first and second formants.

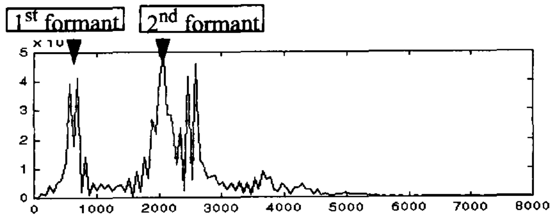

Figure 2 Typical voiced speech spectrum, showing bulges at first and second formant frequencies.

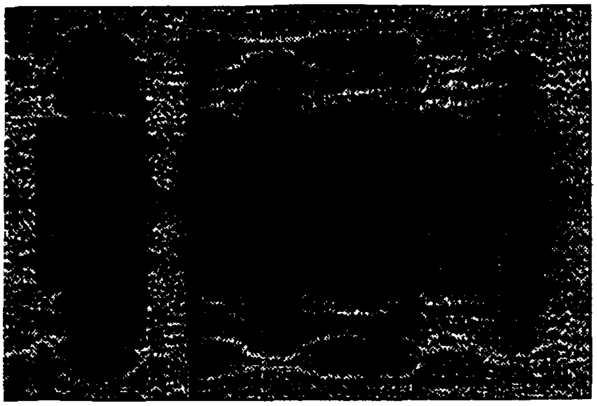

Figure 3: Speech spectrogram computed using autocorrelation based spectrum. Results computed using same analysis parameters as in Fig 3

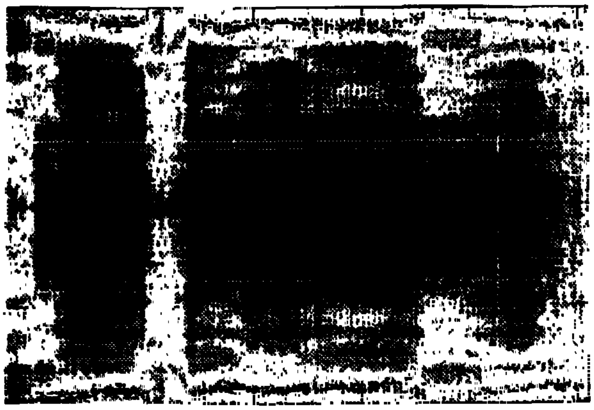

Figure 4: Normal speech spectrogram computed using standard power spectrum

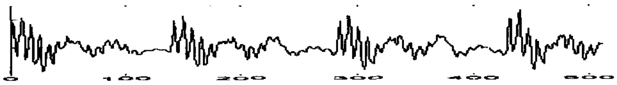

Figure 5 512 randomly selected samples of TIMIT data.

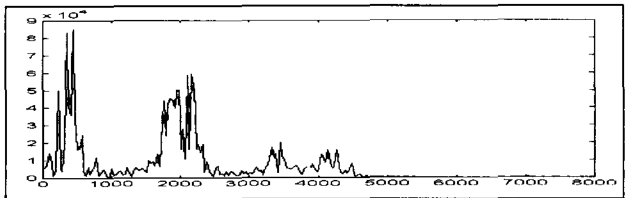

Figure 6 Unmodified narrowband 512 point transform of TIMIT data from Fig. 5, sampled at 16 kHz

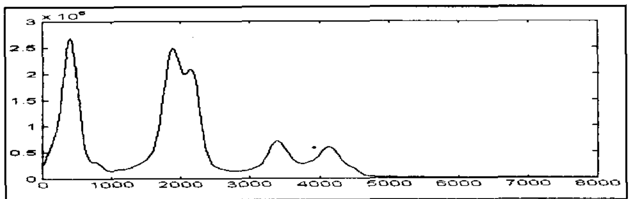

Figure 7 Smoothed narrowband 512 point transform of TIMIT data from Fig. 5.

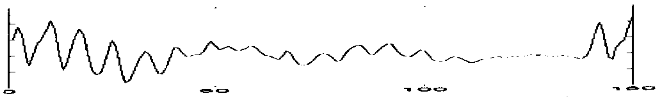

Figure 8 150 samples of randomly selected TIMIT data

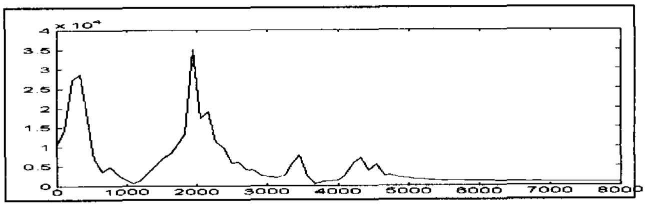

Figure 9 Unmodified sideband 150 point transform of TIMIT data in Fig. 8. Note loss of separation of 2nd and 3rd formants.

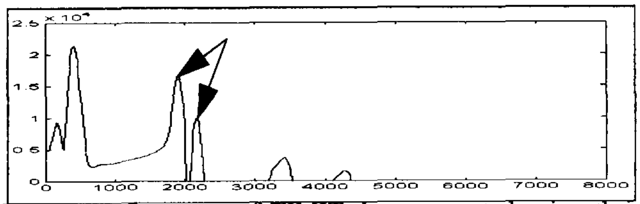

Figure 10 Spectrum of TIMIT data computed using the methods outlined in this paper. A rectified DCT and a 150 point autocorrelation function were applied to the data in Fig. 8. Second and third formants are clearly separated