EXPLOITING TEMPORAL CHANGE OF PITCH IN FORMANT ESTIMATION

Tianyu T. Wang and Thomas F. Quatieri

MIT Lincoln Laboratory
Speech & Hearing Bioscience & Technology Program
Harvard-MIT Division of Health Sciences and Technology
[*ttwang*, *quatieri*]@*ll.mit.edu*

ABSTRACT[1]

This paper considers the problem of obtaining an accurate spectral representation of speech formant structure when the voicing source exhibits a high fundamental frequency. Our work is inspired by auditory perception and physiological modeling studies implicating the use of temporal changes in speech by humans. Specifically, we develop and assess signal processing schemes aimed at exploiting temporal change of pitch as a basis for formant estimation. Our methods are cast in a generalized framework of two-dimensional processing of speech and show quantitative improvements under certain conditions over representations derived from traditional and homomorphic linear prediction. We conclude by highlighting potential benefits of our framework in the particular application of speaker recognition with preliminary results indicating a performance gender-gap closure on subsets of the TIMIT corpus.

Index Terms— formant estimation, source-filter model, effects of pitch, auditory modeling, speaker recognition

1. INTRODUCTION

There is psychophysical evidence suggesting that in speech perception, humans exploit temporal change of pitch, particularly when the stimulus contains a high fundamental frequency (*f0*). In a series of concurrent vowel segregation tasks, McAdams showed that subjects reported an increased prominence for vowels whose *f0* was modulated relative to those that were not modulated [1]. Diehl, et al. showed in vowel perception experiments that a linearly changing *f0* improved subjects' vowel identification accuracy [2]. These effects were greatest when the synthetic source was chosen to have high *f0* (e.g., ~270-400 Hz) such that the speech spectrum is *undersampled*.

In this paper, we aim to improve formant estimation of undersampled speech spectra due to high *f0*. Traditional linear prediction-based (LP) estimation suffers for high *f0* due to aliased autocorrelation estimates [3]. Rahman and Shimamura have suggested the use of homomorphic linear prediction (HLP) for high-pitched formant estimation on single spectral slices, thereby exploiting the source-filter separability characteristics of the cepstrum [4]. In our work, we are instead motivated by evidence that human perception exploits change in pitch as a basis for formant estimation and consider the condition in which *f0* is changing while the underlying formant structure is fixed. For estimation, we propose a two-dimensional (2-D) framework using 2-D transforms of the time-frequency space. In one approach, we project changing spectral harmonics over time to a one-dimensional (1-D) function of frequency. In a second, we draw upon previous work of Quatieri and Ezzat, et al. [5, 6], where localized 2-D transforms of the time-frequency space invoke improved source-vocal tract separation when pitch is changing. Their work has similarities to the physiological auditory modeling efforts of Chi, et al. [7].

This paper is organized as follows. Section 2 reviews the problem of high-pitched formant estimation for undersampled speech spectra in the context of the source-filter production model. Section 3 presents a framework for exploiting temporal change of pitch; specifically, we discuss an auditory-inspired framework of two-dimensional processing of speech as a means to address spectral undersampling. Section 4 gives details of our simulation methods and results for formant frequency estimation. Section 5 gives preliminary results in addressing a male-female performance gap in the particular application of speaker recognition. Section 6 ends with conclusions and future directions.

2. SOURCE-FILTER UNDERSAMPLING

In the source-filter paradigm of speech production, voiced sounds are modeled on short-time scales as the convolution of a periodic impulse train source, $s[n]$, with period T_0 and a linear, time-invariant filter $h[n]$, i.e., $x[n] = s[n]*h[n]$. The speech spectrum, $X(f)$, is then the product of a time-invariant complex spectral envelope $H(f)$ with a harmonic line structure $S(f)$ whose lines are spaced at multiples of the fundamental frequency or "pitch", $f0 = 1/T_0$, i.e., $X(f) = S(f)H(f)$. $X(f)$ can be viewed as the result of uniformly sampling an invariant spectral envelope $H(f)$ at multiples of $f0$ [3]. The ability of $X(f)$ to reflect the underlying formant envelope $H(f)$ depends on the sparseness of $S(f)$. When $f0$ is sufficiently low (e.g., 100 Hz), $H(f)$ can be reasonably traced out by $S(f)$. As $f0$ increases, $S(f)$ becomes more sparse and may miss a formant peak, resulting in a poor representation (i.e., spectral undersampling) of the formant structure (Figure 1).

[1] This work was supported by the Department of Defense under Air Force contract FA8721-05-C-0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. The work of T.T. Wang was additionally supported by the National Institutes of Deafness and Other Communicative Disorders under grant 5 T 32 DC00038.

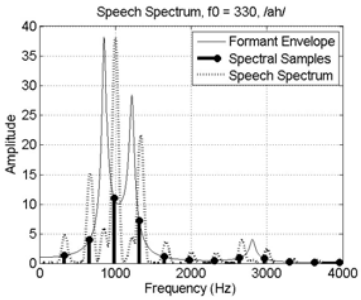

Figure 1. Synthesized vowel /ah/ with 330-Hz pitch. Speech spectrum generated from short-time Fourier analysis with a 20-ms Hamming window.

3. TWO-DIMENSIONAL SPEECH PROCESSING FRAMEWORK

We have observed that spectral undersampling can result in a poor representation of an underlying spectral envelope. Nonetheless, if we consider a condition in which *f0* is changing and the vocal tract remains fixed, the resulting harmonic line structure in a time-frequency space can lead to a more complete view of the vocal tract frequency response. This section presents a 2-D speech processing framework that motivates two approaches for improving spectral estimation.

3.1 Harmonic Projection

Consider the schematic of a short-time Fourier transform (STFT) in Figure 2a showing how changing *f0* provides additional information on a steady vocal tract transfer function. Shaded areas represent the fixed vocal tract frequency response. Under a multiplicative source-filter model, changing harmonics sweep through the spectral envelope over time in a fan-like structure, unlike the horizontal harmonic lines corresponding to a fixed pitch.

Multiple short-time spectral slices computed across time can therefore be viewed as a collection of non-uniform samples of the spectral envelope or alternatively as a projection to the vertical frequency axis as illustrated in Figure 2b. An example of this sampling is shown in Figure 3, contrasting the uniform sampling in Figure 1. It is conceivable that with an appropriate interpolation method, an improved estimate of the underlying spectral envelope can be obtained using this collection relative to that derived from a single spectral slice. One simple method of interpolation is that of averaging short-time spectra across time. For reasons that will become clear later, we think of this alternate transformation as taking the DC value of a 1-D Fourier transform along the time axis at each frequency. In addition, observe that for a given change in *f0* as in Figure 2a and 3, high frequency regions will exhibit broader harmonic sampling relative to lower regions.

3.2 Grating Compression Transform (GCT)

A similarly motivated class of methods makes implicit use of changing pitch for spectral estimation. We adopt the framework proposed by Quatieri and extended by Ezzat, et al. [5, 6]. Consider now a spectrotemporally *localized* region of Figure 2a (rectangle) shown zoomed-in in Figure 2c. Here, the harmonic line structure is roughly parallel and sweeps through a portion of the formant spectral envelope (shaded). The GCT is defined as the short-space Fourier transform computed over such a region [5, 6].

In this work, we model the harmonic line structure as a 2-D sinusoid on a DC pedestal expressed as $1+\cos(\omega_0\Phi(f,t))$ [5]. ω_0 denotes the sinusoidal frequency and $\Phi(t,f)=t\sin\theta+f\cos\theta$ describes its spatial orientation, where θ reflects its skew in the localized region. Denoting the vocal tract contribution as a slowly-varying $A(t,f)$, the complete function over the region in the context of the source-filter spectral model is $S(t,f)=A(t,f)+A(t,f)\cos(\omega_0\Phi(t,f))$. In the GCT, $A(t,f)$ maps to a function concentrated at the origin as represented by the shaded ellipse in Figure 2d. In contrast, the fast-varying $A(t,f)\cos(\omega_0\Phi(t,f))$ maps to a pair of smeared impulses located ω_0 from the GCT origin (measured radially). For changing *f0*, the rotational nature in transforming skewed harmonic lines maps them *off* the scale axis at a non-zero angle θ (represented by solid ellipses, Figure 2d), thereby invoking separability of the source from vocal-tract spectra as can be observed in previous work by Ezzat, et al. [6]. For fixed *f0*, the smeared impulses lie along the scale axis (dotted ellipses, Figure 2d). Note that for fixed *f0*, θ = 0, and source-filter separability can be achieved only if ω_0 is greater than the scale-frequency[2] bandwidth of $A(t,f)$.

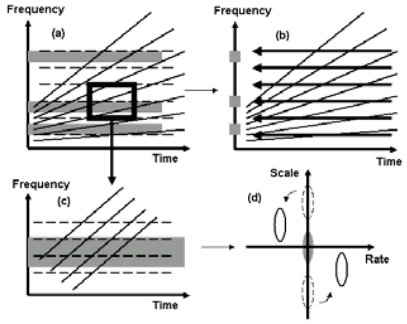

Figure 2. Exploiting changing pitch: (a) Schematic of changing and fixed *f0* across a steady vowel in a STFT; (b) Projection of harmonic samples to a single 1-D frequency axis; (c) Localized spectrotemporal region from (a); (d) Mapping of source-filter speech components in the GCT from (c).

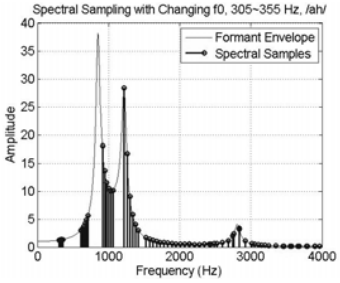

Figure 3. Collection of harmonic samples from pitch sweep ranging from 305~355 Hz. Contrast to *f0* = 330 Hz shown in Figure 1.

For changing *f0*, harmonic lines in the STFT tend to fan out with greatest slope in high-frequency regions. The increased fanning corresponds to larger values of θ when mapped in the GCT. We therefore expect better separability for these regions

[2] Our terminology "rate" and "scale" is borrowed from Chi, et al [7] in referring to modulation content in time and frequency, respectively.

relative to lower frequency regions. This is analogous to the broader harmonic sampling discussed in Section 3.1.

Our model motivates a method spectral estimation distinct from that of a computing a single spectral slice. Specifically, 2-D processing across localized regions of a STFT via low-pass filtering could isolate $A(t,f)$ from $A(t,f)\cos(\omega_0\Phi(t,f))$.

4. SPECTRAL/FORMANT ESTIMATION

To assess the proposed framework, synthesized vowels with changing *f0* were used for spectral estimation and subsequently formant estimation. This section discusses our methodology and presents results of our evaluation.

4.1 Experimental Setup

Pure impulse train source signals with starting *f0* ($f0_s$) ranging from 80-200 Hz for males, 150-350 Hz for females, and 200-450 for children were synthesized with linear pitch increases (*df0*) ranging from 10 to 50 Hz. $f0_s$ and *df0* varied in 5 Hz steps. Source signals were generated at 16 kHz, downsampled to 8 kHz, and filtered with 6th order all-pole models corresponding to the vowels *ah, iy, ey*, *ae*, *oh*, and *oo*; for children, *oh* was replaced by uh because we were unable to find formant data in the literature for *oh*. Formant frequencies (F1, F2, F3) and bandwidths were set to average measurements reported by Stevens [8] and Peterson and Barney [9] for males, females, and children. Vowel durations were set to average measurements on the Switchboard corpus reported in [10].

Based on the framework of Section 3, several techniques were investigated for spectral estimation. All methods resulted in a magnitude spectrum which was fitted to a 6th order all-pole model via LP to obtain formant estimates. These were compared to results of traditional LP (referred to as m = 1) and HLP (m = 2) as proposed in [4]. *f0* estimates were obtained directly from the synthesizer *f0* contour.

4.2 Specific Methods

For m = 1 and 2, a magnitude STFT ($STFT_1$) was computed for each utterance using a 20-ms Hamming window, 1-ms frame interval, and 2048-point discrete-Fourier transform (DFT). The spectral slice located in the middle of the utterance was extracted for LP in m = 1. For m = 2, this slice was also used for cepstral analysis. Specifically, an ideal lifter with cut-off 0.6/*f0* for *f0* < 250 Hz and 0.7/*f0* for *f0* > 250 Hz (as in [4]) was applied to the real cepstrum to generate a spectral estimate for use with LP.

In our first method for exploiting temporal change of *f0* (m = 3), the *f0* contour was used to collect spectral samples across the full duration of each utterance. A STFT was computed with a 1-ms frame interval and an adaptive Blackman window with duration corresponding to three times the *f0* estimate for each frame ($STFT_2$). Since each spectral slice corresponds to a distinct *f0*, peak-picking for the harmonics of *f0* was done using the SEEVOC algorithm to obtain samples of the underlying formant envelope [11]. The resulting collection of values was stored in a single array and linearly interpolated to generate a spectral estimate. To remove confounds of the interpolation method itself, interpolation was also performed on harmonic spectral samples from a *single* spectral slice of $STFT_2$ located in the middle of the utterance (m = 4). As another reference, a spectrum was derived by averaging all spectral slices of $STFT_1$ (m = 5). Note that this is equivalent to extracting values of a GCT (computed from the entire STFT) from its scale axis followed by an inverse Fourier transform.

For m = 6, we implemented filtering of the GCT based on the model presented in Section 3.2 Spectrotemporal regions were extracted from $STFT_1$ with a 2-D Hamming window of time width corresponding to the full duration of the utterance and frequency width of 700 Hz with a 350-Hz frequency shift. Two GCTs were computed for each region using the 2-D DFT. For GCT_1, a 2-D gradient operator was applied to the region to reduce the influence of the DC component followed by the DFT while for GCT_2, the DFT was computed without the gradient. Peak-picking using the max operator on the magnitude of GCT_1 was done to estimate ω_o, and a 2-D elliptic filter (applied to GCT_2) was designed by taking the product of two linear-phase low-pass filters in frequency and time. In time, the pass and stop bands were fixed to $0.25\omega_l$ and $0.5\omega_l$, respectively, where ω_l corresponds to the ω_o estimate derived from the lowest (in frequency) spectrotemporal region of $STFT_1$. In frequency, we used a pass and stop band of $0.5\omega_0$ and ω_o, with ω_o corresponding to the local estimate for each region. Consequently, filtering across different regions was adaptive. A reconstructed time-frequency representation was generated using overlap-add, and a spectral slice was extracted corresponding in time to the middle of the utterance.

4.3 Results

In this work, we used for a goodness metric the average percentage of formant frequency errors across all $f0_s$, *df0*, and vowels. Specifically, for the i^{th} formant and m^{th} method :

$$F_{i,m} = \frac{100}{SDV}\sum_{s=1}^{S}\sum_{d=1}^{D}\sum_{v=1}^{V}\left|\hat{F}_{i,m,s,d,v} - F_i^v\right| / F_i^v$$

with *S, D*, and *V* corresponding to the total number of $f0_s$s, *df0*s, and vowels, respectively. In addition, F_i^v corresponds to the true i^{th} formant frequency for the v^{th} vowel while $\hat{F}_{i,m,s,d,v}$ is its corresponding estimate for the s^{th} starting *f0* and d^{th} *df0*.

Table 1 summarizes the array of methods evaluated. Table 2 gives values of our metric for males, females, and children via traditional LP (m = 1). Tables 3-5 show the relative gains of all other methods (m = 2-6) with respect to this baseline. Our results are consistent with those in [4] suggesting that HLP (m = 2) provides gains over traditional LP (m = 1) in formant estimation, even for high *f0*. Nonetheless, we observe that the method of harmonic projection (m = 3) exhibits the best performance with relative gains up ~84% for F3. Losses incurred by single-slice interpolation (m = 4) are also consistent with the role of changing pitch in improving formant estimation (rather than the interpolation method itself). Spectral slice averaging (m = 5) and filtering in the GCT (m = 6) also provide gains over m = 1 for all formants and over m = 2 for F2 and F3. For m = 6, we believe the smaller gains in F1 stem from the reduced fanning of harmonic lines in lower frequency regions of the STFT; consequently, these are more likely to be mapped *along* the scale axis in the GCT, thereby reducing source-filter separability. Conversely, our framework (m = 3, 5, 6) exhibits the greatest gains for F3, presumably due to broader harmonic sampling for m = 3 and 5 and the increased source-filter separability in the GCT for m = 6 in high-frequency regions (Section 3.1, 3.2). It is curious to note that this finding is consistent with the work of Diehl, et al. [2].

Table 1: Summary of formant estimation methods

m=1	Traditional linear prediction (LP)
m=2	Homomorphic linear prediction (HLP)
m=3	Interpolation of collected harmonic peaks + LP
m=4	Interpolation using single slice + LP
m=5	Time-average of $STFT_1$ + LP
m=6	GCT-based filtering + LP

Table 2. $F_{i,m}$ values for males, female, and children for m = 1.

	males	females	children
i = 1	3.13	4.88	5.68
i = 2	0.86	1.66	2.16
i = 3	0.38	0.77	0.91

Table 3. Relative gains of $F_{i,m}$ for males (%).

	m=2	m=3	m=4	m=5	m=6
i = 1	21.09	62.94	8.95	22.04	10.54
i = 2	27.91	80.23	13.95	59.30	46.51
i = 3	23.68	84.21	-2.63	78.95	55.26

Table 4. Relative gains of $F_{i,m}$ for females (%).

	m=2	m=3	m=4	m=5	m=6
i = 1	17.01	64.96	9.43	9.43	9.22
i = 2	21.69	68.07	7.83	31.93	36.14
i = 3	29.87	80.52	-11.69	57.14	61.04

Table 5. Relative gains of $F_{i,m}$ for children (%).

	m=2	m=3	m=4	m=5	m=6
i = 1	4.93	60.39	12.32	5.99	4.05
i = 2	16.20	50.00	-14.35	19.44	25.93
i = 3	8.79	74.73	-31.87	47.25	47.25

5. Speaker Recognition Experiments on TIMIT

Speaker recognition performance typically exhibits a "gender gap", with better performance on males than females. One contributing factor to this gap may be the poorer spectral representation of formants due to the higher-pitched females. Motivated from our improved spectral estimates (as characterized by improved formant estimation of high-pitch speech), we assessed the value of spectral slice averaging as a basis for addressing this gap on the TIMIT corpus. Nonetheless, we emphasize that spectral undersampling may not be the *only* cause of this gap such that improving formant representation may eliminate other characteristics of the spectrum (e.g., source) that could play a role.

As a baseline feature set, short-time spectra were computed using a 20-ms Hamming window and 10-ms frame interval for use in the mel-cepstrum. Our proposed features use instead a 10-ms Hamming window and a 2-ms frame interval for computing short-time spectra. At each time interval, the average of 5 spectral slices was computed for use in the mel-cepstrum. Both features sets were allied with deltas and used with a 128-mixture-component GMM-UBM back end [13]. Table 6 shows the resulting equal error rates (EER). The proposed features reduce the EER by 2.26% for females while maintaining the performance of males close to the baseline. It appears then, for the TIMIT corpus, that the proposed features close the gender gap.

A caveat in interpreting these results is that the window length and frame interval of the baseline and proposed features are different. To address this confound, we evaluated the performance of females using the same window and frame interval (10 ms and 2 ms) as the proposed features but *without* averaging. This resulted in an EER of 2.85% (2.01% < EER < 3.72%). Though this result shows that the proposed method may afford the gain in females due to the window and frame interval alone, the larger absolute gain in the proposed method appears promising.

Table 6. Comparison of baseline and proposed feature sets on TIMIT males and females. Confidence intervals are at the 95% level.

	Baseline (EER)	Proposed (EER)
Males	1.86% < 2.45% < 3.39%	1.53% < 2.15% < 2.80%
Females	3.12% < 4.41% < 5.64%	1.55% < 2.15% < 3.30%

6. CONCLUSIONS AND FUTURE WORK

In this work, we have shown that spectral estimation methods exploiting changing pitch can improve formant estimation over standard single-frame analysis. Future work will explore alternative representations of the localized spectrotemporal region from Section 3.2 such as applying the logarithm (i.e., $\log[S(t,f)] = \log[A(t,f)] + \log[(1+\cos(\omega_0 \Phi(t,f))]$). A GCT computation then becomes cepstral-like, though better source-filter separability may be obtained than the standard 1-D cepstrum due to the rotational nature of transforming skewed harmonic lines. Because spectral slice averaging showed some gains in speaker recognition, future work will also assess the full range of spectral estimation methods as alternative bases for the mel-cepstrum.

Acknowledgements: The authors thank Douglas Sturim for noting the gender-gap performance issue and Daryush Mehta for use of his vowel-synthesis software.

REFERENCES

[1] S. McAdams, "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am., 86(6): 2148-2159, 1989.
[2] R. Diehl, B. Lindblom, K. Hoemeke, and R.P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phonetics, 24:187-208, 1996.
[3] T. Quatieri, *Discrete-time Speech Signal Processing*, Prentice Hall PTR, Upper Saddle River, NJ, 2002.
[4] M. Rahman, T. Shimamura, "Formant frequency estimation of high-pitched speech by homomorphic prediction," Acoust. Sci. and Tech. Vol. 26, No. 6, pp. 502-510. 2005.
[5] T. Quatieri, "2-D Processing of Speech with Application to Pitch Estimation," Interspeech, Denver, CO 2002.
[6] T. Ezzat, J. Bouvrie, T. Poggio, "Spectro-Temporal Analysis of Speech Using 2-D Gabor Filters," Interspeech, Antwerp, Belgium 2007.
[7] T. Chi, P. Ru, S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds", J Acoust Soc Am., 118(2):887-906. 2005.
[8] K. Stevens, *Acoustic Phonetics*, MIT Press, 1998.
[9] G. Peterson, H. Barney, "Control methods used in a study of the vowels, "J Acoust Soc Am., 24:175-184. 1952.
[10] S. Greenberg, H. Hitchcock, "Stress-Accent and Vowel Quality in the Switchboard Corpus," Workshop on LVCSR, 2001.
[11] D. Paul, "The Spectral Envelope Estimation Vocoder," IEEE Trans.on Acoustics, Speech and Signal Processing, 29:786-794, 1981.
[12] W. Fisher, G. Doddington, K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and Status," Proc. DARPA Workshop on Speech Recognition, Feb. 1986, pp. 93-99.
[13] D. Reynolds, T. Quatieri, R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Dig. Sig. Proc., 10:181-202, 2000.