LONG-TERM FLEXIBLE 2D CEPSTRAL MODELING OF SPEECH SPECTRAL AMPLITUDES

Mohammad Firouzmand [1,2] *& Laurent Girin* [1]

[1] Grenoble Lab. of Images, Speech, Signal & Automatic (GIPSA-lab), INPG, Grenoble, France
[2] Iranian Research Organization for Science & Technology (IROST), Tehran, Iran
{Mohammad.Firouzmand,Laurent.Girin}@gipsa-lab.inpg.fr

Abstract

This paper presents a method for modeling the envelope of spectral amplitude parameters of speech signals in "two dimensions" (2D). It consists of two cascaded modelings: the first one along the frequency axis is the usual cepstrum technique, which consists of modeling the log-scaled spectral envelope with a Discrete Cosine Model (DCM). The second one, along the time axis, consists of modeling the trajectory of the envelope DCM coefficients by another similar DCM model. An iterative algorithm is proposed to optimally fit this 2D-model to the data according to a perceptual criterion based on frequency masking. This approach is shown to provide an efficient and flexible representation of spectral amplitude parameters in terms of coefficient rates, while providing good signal quality, opening new perspectives in very-low bit-rate sinusoidal speech coding.

Index Terms— speech analysis, speech processing, speech coding, speech modeling, speech synthesis.

1. Introduction

The parametric modeling of the spectral envelope of speech signals is a very classical problem in speech processing. It has many applications in speech coding and speech synthesis (including speech transformation). The most well-known techniques are probably the linear predictive coding (LPC) [1] and the discrete cosine model (DCM), which is also known as (discrete/parametric) *cepstrum* modeling [2][3] when applied to log-scaled amplitudes. Since speech signals are non-stationary, such spectral envelopes are generally estimated on a short-term (ST) basis: the length and shift of analysis/synthesis frames are generally of about 10-30ms. In this paper, we deal with the problem of efficiently representing the time-evolution of the spectral envelope on a so-called long term (LT) basis, *i.e.*, from several tens of ms to several hundreds of ms and more. The objective is to provide a compact/sparse and flexible (*i.e.*, with dimensions adapted to local signal characteristics) representation of the 2D smooth spectral envelope. Such representation can be used in speech coding, and in speech transformation systems based on 2D spectral envelope as, e.g., [4].

For this aim, we build on previous works. In [5][6], we proposed to model the LT trajectory of the amplitude of individual speech harmonics using a DCM. An iterative algorithm including perceptual constraints was proposed to jointly estimate the optimal order of the model and its coefficients. This led to a significant reduction of the coefficients rate compared to harmonic ST modeling. In [7], this approach was applied to LSF coefficients that encode the LPC spectrum envelope, using a spectral distortion measure criterion. It was combined with multi-stage vector quantization to provide efficient very-low bit rate LSF coding. A similar approach, using a polynomial model, was proposed by Dusan *et al*. in [8].

In the present paper, we extend the long-term modeling to the cepstrum model: we first model the (log-scaled) envelope of speech spectra by a DCM (Section 2.2). This is done on a usual short-term basis [2][3]. Then we apply a second DCM along the time axis to model the trajectory of the resulting cepstrum coefficients on long sections of speech (Section 2.3). This results in a "2D cepstral model". It is important to note that the present work is not a simple replica of [7] using a cepstrum model instead of the LPC model: In the present work, both the cepstrum order and the temporal model order are variable from one 2D-modeled speech section to the other (whereas the LPC order was fixed to 10 in [7]), and must be estimated. For this aim, we present in Section 2.4 a new iterative algorithm that exploits a perceptual criterion. The proposed method is evaluated in Section 3 in terms of "goodness of fitting", perceptual quality, and coefficient rate.

2. 2D-Cepstrum Modeling

2.1. Analysis

In the present study, we only consider continuously voiced speech sections, considered as (full-band) harmonic signals. The proposed method can be easily applied to sections of unvoiced (or mixed voiced/unvoiced) speech. The signal is thus first segmented into voiced and unvoiced parts, using the fundamental frequency (denoted $\omega_{0,k}$ for frame k) estimation algorithm of Praat [9]. Then, for a given voiced section $s(n)$, running arbitrarily from $n=0$ to N, the successive sets of amplitude parameters to be 2D-modeled are extracted on a short-time basis. We used a classical least mean square (LMS) fitting of the harmonic model (using the $\omega_{0,k}$ measures) with the signal [10]. It provides accurate parameter estimation with very low comput-ational cost. A 25ms analysis window was used, with a hop-size of 20ms, so that the parameter rate is of 50 amplitude vectors per second. This results in K vectors $\mathbf{A}^k = [A_{1,k}\, A_{2,k}\, \ldots\, A_{I_k,k}]^t$, $k = 1$ to K (^{t} denotes the transposed vector/matrix). The vector dimension I_k can vary from one frame to the other, due to the variation of the fundamental frequency.

2.2. Modeling of the spectral envelope

Cepstrum modeling consists of replacing each set of (log-scaled) amplitudes by a sum of cosine functions [2][3]:

$$\hat{A}_k(\omega) = d_{0,k} + 2\sum_{m=1}^{M} d_{m,k} \cos(m\omega) \qquad (1)$$

Here, the cepstrum order, denoted M, is variable from one considered speech section to the other, but it is fixed on each section. This enables: (i) to solve the "variable size problem" of the amplitudes set from one (short-term) analysis frame to the next, (ii) to reduce the size of the parameter set to be time-modeled, since M is generally significantly lower than the size of the amplitude vector; This is a major point for potential application of the proposed method to very low bit-rate speech coding; and (iii) to make frequency transformations such as pitch-scaling easier.

We will see in Section 2.4 how M is estimated for each 2D-modeled section. Given M, each vector $\mathbf{D}^k = [d_{0,k}\ d_{1,k}\ \dots\ d_{M,k}]^t$ of model coefficients is estimated by a LMS fitting of the DCM with the measured (log-scaled) amplitudes, *i.e.*, minimizing:

$$\varepsilon_k = \sum_{i=1}^{I_k} \left| A_{i,k} - \hat{A}_k(i\omega_{0,k}) \right|^2 \qquad (2)$$

If $\mathbf{H}_k$ denotes the $I_k \times (M+1)$ matrix of general entry $h_{i,m} = \cos(mi\omega_{0,k})$ (with a factor 2 for columns with $m>1$), and assuming $M+1 < I_k$, we have:

$$\mathbf{D}^k = (\mathbf{H}_k^t \mathbf{H}_k)^{-1} \mathbf{H}_k^t \mathbf{A}^k \qquad (3)$$

2.3. Modeling the cepstrum coefficients trajectory

Once the spectral envelope modeling has been done for all (short-term) frames of the considered speech section, the second step of the proposed 2D-method is the modeling of the time-trajectory of the envelope DCM coefficients. Thus, these coefficients are now considered along the time axis as $M+1$ sets of K-vectors (as we directly did for amplitudes in [5][6]): $\mathbf{D}_m = [d_{m,1}\ d_{m,2}\ \dots\ d_{m,K}]$, $m = 0$ to M. A second Discrete Cosine Model is then applied on each of these $M+1$ trajectories:

$$\hat{d}_m(n) = c_{m,0} + 2\sum_{p=0}^{P_m} c_{m,p} \cos\left(p\pi \frac{n}{N} \right) \qquad (4)$$

The model order P_m depends here on the rank m of the envelope coefficient. For simplicity, in this study we set the same order $P_m = P$ for all vectors $\mathbf{D}_m$, $m = 0$ to M. The model coefficients $c_{m,p}$ are estimated by minimizing the weighted error:

$$\delta_m = \sum_{k=1}^{K} w_k \left| d_{m,k} - \hat{d}_m(n_k) \right|^2 \qquad (5)$$

where the indexes n_k are the centers of the K analysis frames. Let $\mathbf{M}$ denote the $(P+1) \times K$ matrix of general entry $m_{p,k} = \cos(p\pi n_k/N)$, $\mathbf{W}$ denote the $K \times K$ diagonal matrix that contains the squared weights of (5) on its diagonal (see Section 2.4), and $\mathbf{D}$ denote the matrix that gathers the vectors $\mathbf{D}_m$. Assuming $P+1<K$, the $(M+1) \times (P+1)$ matrix of optimal model coefficients is given by:

$$\mathbf{C} = \mathbf{D}\mathbf{W}\mathbf{M}^t (\mathbf{M}\mathbf{W}\mathbf{M}^t)^{-1} \qquad (6)$$

2.4. 2D-Model orders estimation and fitting algorithm

The evolution of the spectrum envelope can vary widely, for example depending on the length of the considered voiced section, the phoneme sequence, the speaker, the prosody, etc. Thus, we present next the algorithm that is proposed to jointly estimate the orders M and P, the weight matrix in (6), and of course, the 2D-model coefficients $\mathbf{C}$. This algorithm is applied independently to each voiced section of K frames. Note that the proposed 2D modeling can be efficiently exploited in very low bit-rate speech coding if in practice the estimated order P is generally found to be significantly lower than K. The algorithm is split in two parts: The first part consists of tuning M to jointly fit K optimal envelope models to the K amplitude sets according to a mean perceptual criterion. These envelope models are then used in the second part of the algorithm which deals with the time-dimension modeling.

Algorithm for 2D-cepstrum modeling (M is initialized to an arbitrary value, typically 10; R_{1D} and R_{2D} are user-defined ratios lower than 1; typically, $R_{1D} = 0.70$ to 0.90, and $R_{2D} < R_{1D}$)

First part: perceptual cepstrum modeling

1) For $k = 1$ to K, calculate the frequency masking threshold $\mathbf{T}^k = [T(\omega_{0,k})\ \ T(2\omega_{0,k})\ \dots\ T(I_k\omega_{0,k})]^t$ associated with the amplitude vector $\mathbf{A}^k$ by using the model of [11].
2) For $k = 1$ to K, calculate the cepstrum vector $\mathbf{D}^k$ with (3). Calculate the modeling error power function (*square* denotes the entry-wise square function):
$$f(\mathbf{E}^k) = \frac{1}{2} square(\mathbf{A}^k - \mathbf{H}_k \mathbf{D}^k)$$
Calculate the percentage R_k of positive entries in $\mathbf{T}^k - f(\mathbf{E}^k)$ (*i.e.*, the percentage of harmonics correctly modeled according to the perceptual criterion).
3) Calculate R_{mean} the mean value of R_k across k. If $R_{mean} \geq R_{1D}$, M is decreased by 1, else M is increased by 1. Then return to step 2 until the minimum value of M for which $R_{mean} \geq R_{1D}$ is found and selected.

Second part: time-modeling of the cepstrum coefficients

4) Initiate P to an arbitrary value significantly lower than K, e.g., set P to the entire part of $K/4$. Calculate the weight matrix of (6) by $w_k = 1/SD_k$, with:
$$SD_k = \sqrt{\frac{1}{I_k} \sum_{i=1}^{I_k} \left[20\log_{10} A_{i,k} - 20\log_{10} \hat{A}_{i,k} \right]^2}$$
5) Calculate the model matrix $\mathbf{C}$ with (6).
6) Decode the K envelope models with: $\tilde{\mathbf{D}} = \mathbf{C}\mathbf{M}$. Let $\tilde{\mathbf{D}}^k$ denote the k-th column of $\tilde{\mathbf{D}}$.
7) For each time index $k = 1$ to K, decode the 2D-modeled amplitudes with: $\tilde{\mathbf{A}}^k = \mathbf{H}_k \tilde{\mathbf{D}}^k$.
8) Calculate the new value of the ratio R_{mean} of step 2 after replacing the amplitudes modeled in the first part of the algorithm with the 2D-modeled amplitudes of step 7. If $R_{mean} \geq R_{2D}$, P is decreased by 1, else P is increased by 1. Then return to step 5 until the minimum value of P for which $R_{mean} \geq R_{2D}$ is found and selected.

Note that the time weights in Step 4 are used to take into account the relative accuracy of the cepstrum models across frames. The spectral distortion SD_k is very usual in speech processing [12]. The more accurate is the envelope model on a given frame, the lower is the corresponding spectral distortion, and the larger is the weight of that frame in the time-modeling process. This weighting process has been shown to generally provide a better global fitting of the 2D-cepstrum compared to if no weights are used.

2.5. Synthesis

For the evaluation, synthesis signals were generated from the 2D-modeled amplitudes. Usual frame-to-frame linear interpolation of the amplitudes was used, as well as interpolation of frequencies and phase measured parameters, followed by an application of the equations of the sinusoidal model of speech, as in [13]. This step

includes a "birth and death" process (*i.e.*, interpolation from or towards zero) for components that go above the Nyquist frequency. Note that a low-bit rate speech coder that would use the proposed method would include the coding of the fundamental frequency trajectory which is used in Step 7 (in matrix $\mathbf{H}_k$). In this paper, the modeling process only concerns the voiced parts of speech. Thus, the unvoiced sections are unchanged and concatenated with the 2D-modeled voiced sections with local overlap-add windowing to avoid audible artifacts [10].

3. Experiments

To test the presented method, we used 8 kHz signals of continuous speech produced by 12 different speakers (6 male speakers and 6 female speakers). About 3500 voiced segments of different sizes were modeled, representing more than 13 minutes of speech.

3.1. General observations and modeling accuracy

First, for suitable R_{2D} values (say around 0.75), the algorithm was shown to adapt correctly to the different "shapes" of the modeled speech sections. For example, along the frequency axis, the order M was shown to vary from low values (e.g., 4) for spectra with poor relief, to usual values for female speech coding (e.g., 10-11) and male speech coding (e.g., 15-16, see [1]), and more for "rich" spectra. Along the time axis, the order P also varied a lot, depending on the length and the content of the 2D-modeled section. It was found to be generally significantly lower than K. The modeled trajectories of the cepstrum coefficients are thus smoothed versions of their raw trajectories, since the model is composed of smooth cosine functions (see Fig. 1). Yet, the 2D-modeling was shown to provide amplitudes trajectories that are close to the measured amplitudes. This is guaranteed by the perceptual constraint that guides the behavior of the fitting algorithm: At the end of the algorithm, R_{2D} percent of the 2D-modeled amplitudes are assumed to fulfill the perceptual constraint (*i.e.*, the modeling error is below the masking threshold model, and is thus expected to be inaudible). By choosing the settings of R_{1D} and R_{2D}, one can balance the modeling accuracy between the frequency and time dimensions. Different values of R_{1D} for the same R_{2D} can lead to quite different results. A good balance is generally obtained with $R_{2D} \approx 0.85 \times R_{1D}$. In the next section, we give a quantitative assessment of this point.

3.2. Coefficient rates

In this sub-section, we report average rates for the coefficients of the 2D-model (*i.e.*, the coefficients of **C** which is the information to be transmitted to encode the spectral amplitudes with the 2D-technique). Table 1 provides the results that were obtained on the entire test database by varying both R_{1D} and R_{2D}, for both female and male speech. Optimal rates, defined as the lower values obtained for each fixed R_{2D} value are marked in grey cells. It can be seen that these optimal values are regularly distributed (on an inner diagonal). This shows that a reasonable difference between R_{2D} and R_{1D} must be assumed. Indeed, we observed that a too low value of R_{1D} provides poor spectral modeling which cannot be compensated by an accurate time-modeling of the resulting coefficients. On the other hand, for a given R_{2D}, increasing the accuracy of the envelope modeling by increasing R_{1D} beyond the optimal value leads to increase the coefficient rate in a useless manner. Finally, a 10% difference (or again, an about 0.85 ratio) between R_{2D} and R_{1D} seems appropriate.

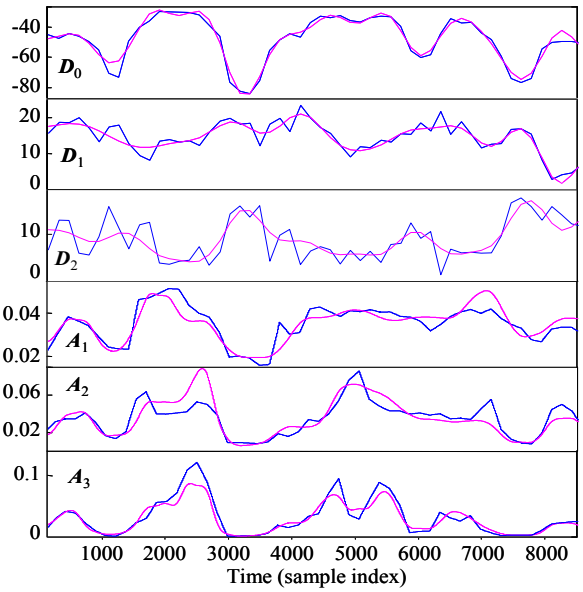

Figure 1: 2D cepstrum modeling for a long voiced section (1.1s of male speech at 8 kHz, K=55). Three first figures: time-trajectories of the 3 first cepstrum coefficients: original coefficients (*i.e.*, from (3); raw blue lines) and modeled coefficients (smooth magenta line); Three last figures: amplitude trajectories of the 3 first harmonics: original measures (raw blue lines) and 2D-modeled amplitudes (smooth magenta lines). $R_{1D} = 0.90$, $R_{2D} = 0.70$, $M = 13$, $P = 19$.

R_{1D} \ R_{2D}	80%	75%	70%	65%	60%	80%	75%	70%	65%	60%
90%	345	296	248	207	175	422	384	323	279	244
85%	361	266	222	183	153	462	321	278	234	203
80%	*	273	205	167	137	*	338	246	205	174
75%	*	*	208	158	126	*	*	254	189	157
70%	*	*	*	159	120	*	*	*	192	141
65%	*	*	*	*	121	*	*	*	*	146
1D	395	341	297	260	230	462	395	334	280	243
Gain	12.7	22.0	30.1	39.2	47.8	8.7	18.7	26.3	32.5	42.0

Table 1: Coefficient rates of the amplitude 2D-model for the female voices (on the left; averaged over 1800 speech segments) and for the male voices (on the right; averaged over 1700 speech segments), and for different {R_{1D} ; R_{2D}} configurations. "1D" stands for the 1D-model used as a reference (see the text), and "Gain" stands for the relative gain between 2D- and 1D-method (in %).

The optimal coefficient rates are quite low. For comparison, we also calculated the rates provided by the "1D" usual approach (*i.e.*, only the short-term cepstrum modeling is considered, and resulting coefficients are transmitted every 20ms). For this 1D-method, we set R_{1D} to the 2D-method target value R_{2D} for fair comparison (this way, both 1D- and 2D-method provide the same final percentage of amplitudes correctly modeled according to the perceptual criterion). As can been seen from Table 1, the 2D modeling method provides quite large gains in coefficients rate compared to the 1D modeling method. For female speech, gains ranging from 12.7% to 47.8% are obtained, depending on the configuration. For male speech, the gains are within 8.7%–42.0% (note that, as is usual in speech modeling, all rates are higher for male speech than for female speech). Thus, the 2D-modeling strategy leads to significantly decrease the rates of model coefficients (until about 50% for the lower constraints).

3.3. Signal quality

Informal extensive listening tests reveal that, for values of R_{2D} about 0.75 (and correct setting of R_{1D}), the synthesized signals are of good quality, fairly close to the originals (see Fig. 2 for a visual illustration), and very close to the signals synthesized from the measured amplitude parameters, without any modeling. Also, they were found to be very close to the signals synthesized from the 1D-modeled parameters, with the same overall target ratio (*i.e.*, R_{1D} of the 1D-method = R_{2D} of the 2D-method), while the coefficient rates are significantly lower for the 2D-method. If R_{2D} is between 0.70 and 0.60, the quality is lowered, but remains quite fair. For lower R_{2D}, say below 0.60, the trajectories of the DCM coefficients tends to be "over-simplified" (since the order P gets lower), and so are the trajectories of the envelope spectrum. Thus, the resulting signal, although good-sounding, moves away from the original signal, tending towards some "hypo-articulated" utterance.

To confirm the efficiency of the proposed method, two formal listening tests were conducted, in a quiet environment, using a high-quality PC soundcard and Sennheiser HD202 headphones. Twelve sentences were used, from 6 speakers (3 male, 3 female, 2 sentences from each speaker). The settings were: R_{1D} = 80% and R_{2D} = 70% (hence R_{1D} = 70% for the 1D-modeled signals). The average gain in coefficient rate of the 2D- over the 1D-method was 30% for female voices and 26% for male voices (hence the selected signals represent well the general results of Table 1). Ten naive listeners with normal hearing were first asked to choose the signal with the best perceived quality among the pairs of 1D/2D-modeled signals, presented in random order (*i.e.* perform an AB test). In the second test, the signal synthesized from the measured amplitudes (without any modeling) was provided, and the listeners were asked to point out which of the 1D/2D-modeled signal was closer to this reference signal (*i.e.*, perform an XAB test).

For the AB test, the overall preference score across sentences and subjects is 70.8% for the 2D-method vs. 29.2% for the 1D-method. Therefore, this test reveals a strong preference for the 2D-model over the 1D-model. This is confirmed by the XAB test: a score of 79.2% vs. 20.8% in favor of the 2D-modeled signals shows that they are clearly judged more faithful to the reference signals (note that a training effect may have occurred since the XAB test has been done systematically after the AB test). Given that the 2D- and 1D-modeled signals were generally found to be very close in the informal tests, we were (positively) surprised by the extent of the score unbalances. These results clearly validate the 2D proposed model from a perceptual point of view.

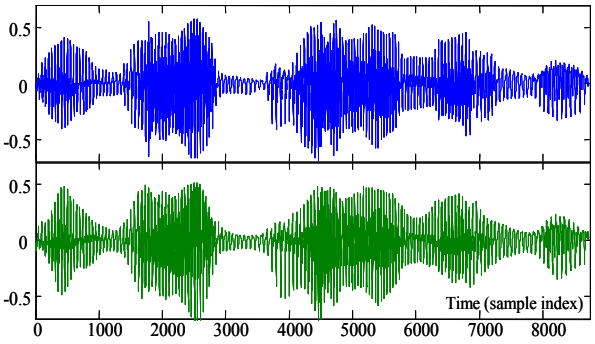

Figure 2: Top: original signal (same signal as the one used for Fig. 1); Bottom: signal synthesized with the 2D-modeled amplitudes (R_{2D} = 0.70 ; R_{1D} = 0.90, M = 13, P = 19, K = 55).

4. Conclusion

This work has confirmed the robustness and generality of the DCM, which is adequate for modeling both spectral envelopes (as shown before in [2][3]) and time-trajectories of model parameters (as shown before in [5][6][7]). Plugging these two aspects together in a 2D cepstrum modeling scheme has led to further advances. Specifically, in this study, the main reason for the efficiency of the 2D-modeling is the intrinsic double variable-rate property: both spectral and time model orders M and P are automatically adjusted to the local signal characteristics. Such flexible 2D-approach can lead to significantly reduce the number of model coefficients for spectral amplitudes representation, while preserving a very good quality for the synthesized signals, as revealed by listening tests. This opens new doors to very-low bit-rate sinusoidal coding of speech/audio signals.

Future works will concern i) the adaptation of the method to unvoiced/mixed V-UV speech sections, ii) quantization issues, and iii) the use of the method in a sinusoidal speech coder at very low bit-rate and with large delay tolerance. For such application, phase information can be reduced to the fundamental frequency trajectory, which can also be long-term modeled [6].

5. References

1. J.D. Markel & A.H.Jr. Gray, *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.
2. O. Cappé, J. Laroche & E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," *Proc. IEEE Work. Applic. Signal Proc. Audio Acoustics*, New Paltz, 1995.
3. T. Galas & X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals," *Proc. Int. Computer Music Conference,* Glasgow, 1990, pp. 82-84.
4. H. Kawahara, I. Masuda-Katsuse & A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction", *Speech Com.*, **27**(3/4), pp. 187-207, 1999.
5. M. Firouzmand & L. Girin, "Perceptually weighted long-term modeling of sinusoidal speech amplitude trajectories," *Proc. IEEE ICASSP 2005,* Philadelphia, USA, 2005.
6. L. Girin, M. Firouzmand & S. Marchand, "Perceptual long-term variable-rate sinusoidal modeling of speech," *IEEE Trans. Speech and Audio Proc.*, **15**(3), pp. 851-861, 2007.
7. L. Girin, Long-term quantization of LSF parameters, *Proc. IEEE ICASSP 2007*, Honolulu, Hawaii, USA, 2007.
8. S. Dusan, J. Flanagan, A. Karve, and M. Balaraman "Speech compression by polynomial approximation," *IEEE Trans. Audio Speech, and Language Proc.,* **15**(2), pp. 387-395, 2007.
9. P. Boersma and D. Weenink, *Praat: doing phonetics by computer,* software available at http://www.praat.org/.
10. E. B. George & M. J. T. Smith, "Speech analysis, synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Proc.,* **5**(5), pp. 389-406, 1997.
11. ISO/IEC JTC1/SC29/WG11 MPEG IS11172-3 Information technology – Coding of moving pictures and associated audio for digital storage at up to about 1.5 Mbits/s, Part 3: Audio, 1992.
12. R. M. Gray, A. Buzo, A. H. Gray and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust. Speech and Signal Proc.*, **28**(4), pp. 367-376, 1980.
13. R. J. McAulay & T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech and Signal Proc.*, **34**(4), 1986, pp. 744-754.