SPEECH RECONSTRUCTION FROM MEL FREQUENCY CEPSTRAL COEFFICIENTS AND PITCH FREQUENCY

Dan Chazan, Ron Hoory, Gilad Cohen and Meir Zibulski

IBM Research Laboratory in Haifa
MATAM Haifa 31905, ISRAEL
chazan@il.ibm.com

ABSTRACT

This paper presents a novel low complexity, frequency domain algorithm for reconstruction of speech from the mel-frequency cepstral coefficients (MFCC), commonly used by speech recognition systems, and the pitch frequency values. The reconstruction technique is based on the sinusoidal speech representation. A set of sine-wave frequencies is derived using the pitch frequency and voicing decisions, and synthetic phases are then assigned to each respective sine wave. The sine-wave amplitudes are generated by sampling a linear combination of frequency domain basis functions. The basis function gains are determined such that the mel-frequency binned spectrum of the reconstructed speech is similar to the mel-frequency binned spectrum, obtained from the original MFCC vector by IDCT and antilog operations. Natural sounding, good quality intelligible speech is obtained by this procedure.

1. INTRODUCTION

The mel-frequency cepstral coefficients (MFCC) [1] are commonly used in many speech recognition systems [2]. They play the role of the acoustic feature vectors, extracted by the front-end of the speech recognizer. The pitch frequency may also be used for recognition, especially for tonal languages (e.g. Mandarin Chinese).

Speech recognition technologies are finding their way into client-server applications. In some scenarios, a client compresses the speech during recording using one of the common speech compression techniques. The compressed speech is then stored on the client for future uploading to a speech recognition server ("*deferred recognition*"), or transmitted over a bandwidth limited channel to a server for real time speech recognition. In addition, the client may also support playback of the compressed speech. However, compressing the speech waveform usually increases recognition errors (especially for dictation tasks), and therefore high quality compression methods should be used. These methods require large processing power or high bit rates, which are often not available in client-server systems (e.g., when a *thin-client* is used).

An alternative to the above would be to perform the MFCC feature vectors extraction and compression on the client. The compressed feature vectors can then be stored locally or transmitted to the speech recognition server. It has been shown that MFCC compression at a bit rate of 4.0 Kbps [3] is possible for continuous speech recognition tasks, without impairing recognition rates. If the pitch frequency is also determined and compressed during recording, such a client can reconstruct the speech for playback purposes, using the algorithm presented in this paper. This alleviates the need for a separate compression scheme to be active during recording. In addition, playback can be performed on the server, without the need to transmit information other than the required MFCC feature vector and the pitch.

Techniques for speech reconstruction from mel-cepstral or cepstral like parameters have been previously proposed [4], [5]. However, in both cases, the definition of the cepstral parameters is rather specific, chosen a priori to allow spectral reconstruction and not in line with the MFCC features used in speech recognition systems.

The MFCC feature vector is computed by integration of the spectrum within triangular *bins* arranged on a mel-frequency axis to form a mel frequency *binned spectrum*, followed by a log and DCT operations. The MFCC vectors and pitch if used, contain sufficient information for continuous speech recognition (at various rates of error). However, it is not obvious that good quality speech can be regenerated from them, since during the feature extraction, a significant amount of information is lost. This includes the phase information, discarded during the spectrum calculation (absolute value of the windowed DFT) and the fine details of the spectrum discarded during the integration. In addition, in most cases some of the higher order cepstral coefficients are discarded. This information is considered "redundant" to some extent for the recognition process.

The proposed reconstruction algorithm synthesizes the extra information required for speech reproduction. The reconstructed speech will have the same pitch as the original speech and a similar binned spectrum (or MFCC).

Section 2 presents the reconstruction scheme. Sections 3 and 4 describe the computational details of the main algorithmic blocks. Results and future directions are discussed in section 5.

2. RECONSTRUCTION ALGORITHM

The algorithm is based on using a sinusoidal model [6], [7], where a short-term (ST) speech signal is represented by a sum of sine waves. The ST-signal is characterized by the frequencies, amplitudes and phases of the sine-wave components. Given the MFCC vector, the pitch frequency and

the voicing decision, a set of sine-wave frequencies are determined, according to the required synthesis sampling rate. Synthetic phases are then generated and assigned to them. The sine-wave amplitudes are estimated according to the desired binned spectrum, calculated from the given MFCC vector. Finally, the short time Fourier transform (STFT) is reconstructed and converted to a time domain signal by an overlap-add method.

The reconstruction scheme is presented in Fig. 1. It is applied every 10 msec frame using an MFCC feature vector, a voicing decision and a pitch frequency value for the voiced case. The output is a windowed short-term signal (20 msec long). After an overlap-add with the ST-signal of previous frames, a 10 msec speech signal frame is produced.

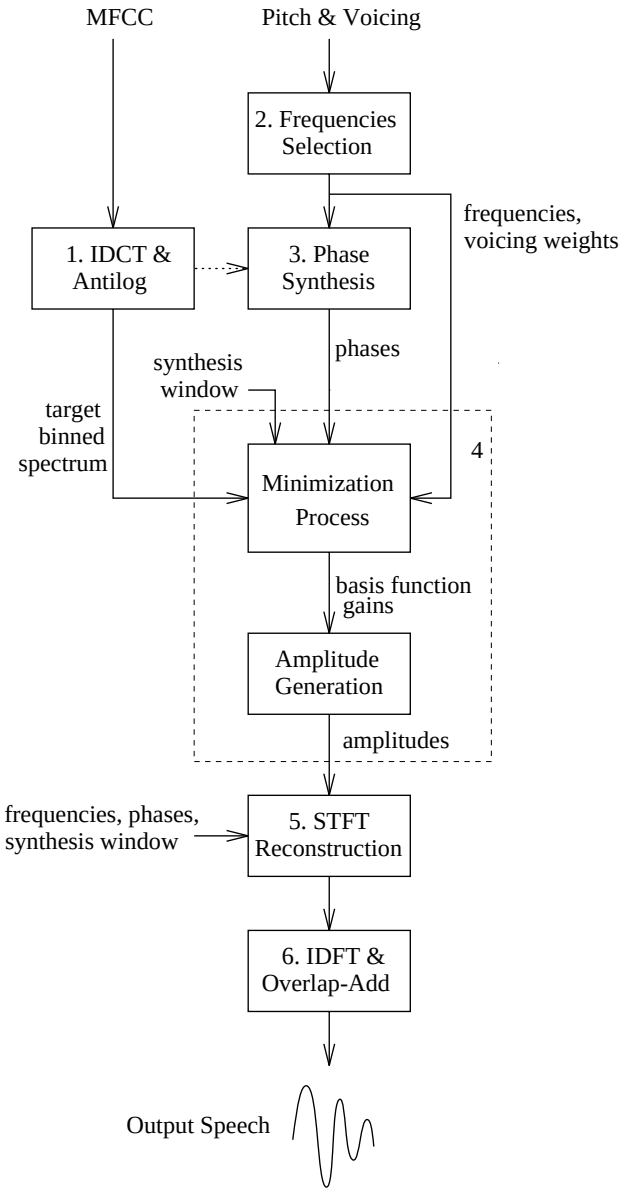

Figure 1: Block diagram of the reconstruction scheme.

The algorithm is comprised of the following stages:

1. **Mel-cepstrum to binned spectrum conversion** (block 1 in Fig. 1) – This is the mathematical inversion of the log and DCT operations carried out during the MFCC computation. If the number of mel-cepstral coefficients is smaller than the number of frequency bins, the MFCC vector is expanded by adding zero coefficients.
2. **Frequencies and voicing weights selection** (block 2) – At this point a set of sine-wave frequencies is selected according to the voicing decision and pitch frequency. Each sine-wave component is assigned a voicing weight according to a predetermined template representing the voicing degree. This is described in subsection 3.1.
3. **Phase synthesis** (block 3) – In this stage the sine-wave phases corresponding to the sine-wave frequencies are synthesized. The phase synthesis algorithm is described in subsection 3.2.
4. **Amplitude generation** (blocks 4) – The sine-wave amplitudes are calculated in accordance with the target binned spectrum, using a set of frequency domain basis functions. The algorithm is described in section 4.
5. **STFT reconstruction** (block 5) – The frequencies, phases and amplitudes are combined to form a sine-wave representation. The final reconstructed STFT is constructed from the sine waves by a convolution procedure.
6. **Conversion to time domain** (block 6) – This is the final stage in which the time domain signal frame is computed. A windowed short-term signal is obtained by an IDFT and then overlap-added to the previous ST signal.

An additional step may be necessary in the time domain if the original MFCC vectors were computed on a pre-emphasized version of the speech signal, as is commonly done in speech recognition systems [2]. A de-emphasis filter, which inverts the effect of the pre-emphasis filter should then be applied.

3. HARMONIC EXCITATION GENERATION

For every short-term analysis interval, the sine-wave speech representation is defined as the set $\{f_i, A_i, \varphi_i\}_{i=0}^{N-1}$, where N is the number of sine-wave components, and f_i, A_i and φ_i are their frequencies, amplitudes and phases respectively. The "harmonic excitation" is defined by the sine-wave frequencies and phases, and the amplitudes are associated with the actual spectral shaping.

3.1. Frequencies and voicing weights selection

This step consists of selecting the N sine-wave frequencies $\{f_i\}_{i=0}^{N-1}$ and the voicing weights $\{\rho_i\}_{i=0}^{N-1}$. The voicing weights are assigned values between 0 and 1, and are embedded in the amplitude calculation (see definition (4) below).

In voiced frames, the frequencies are chosen as the pitch harmonics, $f_i = i \cdot F_0$ where F_0 is the pitch frequency. To reconstruct a more natural sound, additional "unvoiced" frequencies are arbitrarily added. These are mostly concentrated in the upper frequency band. The pitch harmonics are assigned voicing weights of 1. The additional unvoiced frequencies are assigned weights rising from 0 to 1 as the frequency increases.

In unvoiced frames, the frequencies selected correspond to the frequencies of the DFT points, each assigned an equal weight of 1.

3.2. Phase synthesis

The unvoiced frequencies are each assigned a uniformly random phase, i.e., $\varphi_i = U[-\pi, \pi]$. For the voiced frequencies (pitch harmonics), a phase generation scheme based on [7] is used. In this case voiced phases may consist of two components:

$$\varphi_i = i\phi_0(kT) + \Phi_i, \quad i = 0, 1, \ldots, N-1,$$

where k is the frame index, T is the frame duration, ϕ_0 is the pitch excitation phase and Φ_i is an optional phase offset associated with the vocal tract. The pitch excitation phase is calculated such that the pitch harmonics are continuous over frame boundaries. It is evaluated using the following expression:

$$\begin{aligned} \phi_0(kT) &= \int_0^{kT} 2\pi F_0(\sigma) d\sigma \\ &= \phi_0[(k-1)T] + 2\pi \left(F_0^{(k-1)} + F_0^{(k)} \right) \frac{T}{2}, \end{aligned}$$

where $F_0(\cdot)$ is the continuous-time pitch frequency function and $F_0^{(k)}$ is the pitch at frame k. Linear interpolation of the pitch frequency function is assumed between the frames.

The additional phase Φ_i can be derived using a minimum phase model. Assuming the vocal tract transfer function is minimum phase, a relation between its amplitude and phase can be expressed via the complex cepstrum [8]. The input binned spectrum may be used in order to estimate a smoothed version of the speech spectrum, and from it the additional phase components Φ_i can be calculated at the frequencies f_i.

4. AMPLITUDE GENERATION

4.1. Minimization problem definition and solution

The objective of this stage is to reconstruct the sequence $\{A_i\}_{i=0}^{N-1}$, given the sine-wave frequencies $\{f_i\}_{i=0}^{N-1}$ and the target binned spectrum $\{\beta_k\}_{k=0}^{M-1}$, where M is the number of frequency bins (typically 24).

In general, given a sine-wave speech representation, the discrete-time Fourier transform[1] of the corresponding windowed short-term signal is given by:

$$S(f) = \sum_{i=0}^{N-1} A_i e^{j\varphi_i} W(f - f_i), \tag{1}$$

where $W(f)$ is the Fourier transform of the time domain window (typically a Hamming window). $W(f)$ is real, since the window is symmetric, and for all practical purposes it may be assumed to have a zero value for $|f| > BW_{win}$.

The binned spectrum of $S(f)$ is computed using the same filter-bank as in the MFCC computation [1]. The k'th component of the binned spectrum is computed by:

$$B_k = \sum_l H_k(l\Delta f) \, |S(l\Delta f)|, \quad k = 0, 1, \ldots, M-1, \tag{2}$$

where $H_k(f)$ is the frequency response of the k'th triangular bandpass filter, Δf is the frequency resolution of the DFT utilized, and the summation is over all DFT points. By substituting (1) in (2) we obtain the reconstructed binned spectrum as a function of the sine-wave amplitudes A_i:

$$B_k = \sum_l H_k(l\Delta f) \left| \sum_{i=0}^{N-1} A_i e^{j\varphi_i} W(l\Delta f - f_i) \right|. \tag{3}$$

Given the frequencies f_i and target binned spectrum β_k, we wish to find $\{A_i\}_{i=0}^{N-1}$ s.t. $\sum_{k=0}^{M-1} [B_k - \beta_k]^2$ is minimized and $A_i \geq 0$, $\forall i$. This minimization problem is underdetermined (usually $N > M$). In order to decrease the number of degrees of freedom, additional constraints should be introduced. We define a set of M positive frequency domain basis functions: $\psi_m(f)$, $m = 0, \ldots, M-1$, and define the set of amplitudes to be samples of a linear combination of these basis functions at the sine-wave frequencies f_i:

$$A_i \triangleq \rho_i \sum_{m=0}^{M-1} x_m \psi_m(f_i), \quad i = 0, 1, \ldots, N-1, \tag{4}$$

where x_m are the non-negative basis function gains ($x_m \geq 0$, $\forall m$) and ρ_i are the respective voicing weights as described in subsection 3.1. By substituting (4) in (3) and changing the order of summation, we obtain an expression for the binned spectrum of the reconstructed signal as a function of the basis function gains x_m:

$$B_k = \sum_l H_k(l\Delta f) \cdot \left| \sum_{m=0}^{M-1} x_m S^{(m)}(l\Delta f) \right|, \tag{5}$$

where:

$$S^{(m)}(f) = \sum_{i=0}^{N-1} \psi_m(f_i) \rho_i e^{j\varphi_i} W(f - f_i) \ , \ m = 0, 1, \ldots, M-1$$

is the discrete-time Fourier transform of the signal reconstructed using the m'th basis function alone. Note that $S(f) = \sum_{m=0}^{M-1} x_m S^{(m)}(f)$.

The expression (5) is replaced by an approximation $\tilde{B}_k$, in which the absolute of sums is replaced by the sum of absolutes:

$$\tilde{B}_k = \sum_{m=0}^{M-1} x_m \sum_l H_k(l\Delta f) \cdot \left| S^{(m)}(l\Delta f) \right|, \tag{6}$$

[1] The discrete-time Fourier transform of a signal $x[n]$ is defined as: $X(f) \triangleq \sum_n x[n] \exp(-jn2\pi f/f_s)$, where f_s is the sampling frequency.

where $k = 0, 1, \ldots, M-1$. The resulting approximated binned spectrum is linear in respect to the basis function gains. The approximation (6) is accurate when there is no overlap between the supports of the summed terms in (1), i.e., $F_0 > 2BW_{win}$, or when the overlap is small. Moreover, it can be shown that (6) is a good approximation for particular selections of the basis functions. Define:

$$R_{k,m} \triangleq \sum_l H_k(l\Delta f) \cdot \left| S^{(m)}(l\Delta f) \right|.$$

Then (6) becomes: $\tilde{B}_k = \sum_{m=0}^{M-1} x_m R_{k,m}$.

We wish to find a set of non-negative basis function gains, s.t.

$$\sum_{k=0}^{M-1} \left[\sum_{m=0}^{M-1} x_m R_{k,m} - \beta_k \right]^2$$

is minimized. Introducing an $M \times M$ matrix $\mathbf{R}$, whose entries are $R_{k,m}$, and vectors $\mathbf{x}, \beta$, which are $M \times 1$ column vectors, the above minimization problem can conveniently be written in a matrix form:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \geq 0} \|\mathbf{R}\mathbf{x} - \beta\|^2, \qquad (7)$$

where $\hat{\mathbf{x}}$ is the optimal basis function gain vector.

The least squares problem (7) can be solved by any number of iterative techniques (e.g. steepest decent) [9], under the constraint of a non-negative solution. In practice, a simple iterative method, where at each iteration a different basis function gain is updated separately, is used to minimize (7).

Finally, the amplitudes A_i are reproduced using (4). The STFT can then be reconstructed from the sine-wave frequencies, amplitudes and phases using (1).

4.2. Basis function selection

The choice of basis functions $\psi_m(f)$ is critical for obtaining a good reconstruction. The space spanned by the M basis functions represents the structure of the spectral envelope. It was found that basis functions similar in nature to $H_k(f)$, i.e., with the same supports and center frequencies, give good reconstruction results. A typical set of basis functions is shown in Fig. 2. Such basis functions assure that the reconstruction accuracy increases in lower frequencies, where the ear is more sensitive, and maintains the spectral envelope formant peaks. Because of the narrow supports, the matrix $\mathbf{R}$ becomes sparse (banded) and the minimization problem (7) may be solved efficiently.

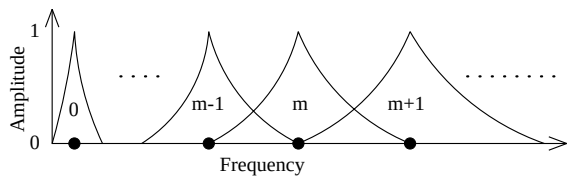

Figure 2: Typical 50% overlapping basis functions $\psi_m(f)$.

5. RESULTS AND DISCUSSION

Speech signals at sampling rates in the range 8-16 kHz were reconstructed using the described algorithm. It was observed that in spite of approximation (6), the actual binned spectra computed after reconstruction were indeed very close to the target binned spectra. Natural sounding, good quality intelligible speech was achieved, which is applicable for playback in Voice Recognition systems.

Reconstruction quality is highly dependent on the accuracy of the pitch estimation. MFCC feature vector dimensions of 13 and 24 were used (with $M = 24$ frequency bins), resulting in a noticeable improvement in quality when using a vector dimension of 24. Using the additional minimum phase as described in section 3.2 resulted in minor quality improvement and may be omitted due to complexity considerations. Reconstruction was demonstrated at 4%-6% real time on a Pentium II 266 MHz PC workstation (depending on sampling rate and MFCC dimension).

Work is in progress to improve the reconstructed audio quality. This includes psychoacoustic post filtering, weighted least square solutions in (7), improved mixture of voiced and unvoiced frequencies, etc. An improved feature extraction algorithm will include pitch frequency and subband voicing degree indicators extracted in a closed loop manner (analysis-by-synthesis), which may enhance speech quality at the expense of higher processing power.

6. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.

[2] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, pp. 45–57, Sept. 1996.

[3] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *Proceedings of ICASSP*, 1998.

[4] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "Celp coding based on mel cepstral analysis," in *Proceedings of ICASSP*, p. 33, 1995.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, p. 137, Mar. 1998.

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

[7] R. J. McAulay and T. F. Quatiery, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 4, pp. 121–170, Elsevier, 1995.

[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[9] Å. Björck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.