Speech Production Based on the Mel-Frequency Cepstral Coefficients

Zbynì k Tychtl and *Josef Psutka*

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeò, Czech Republic

tychtl@kky.zcu.cz, psutka@kky.zcu.cz

ABSTRACT

The mel-frequency cepstral coefficients (MFCCs) are frequently used as a speech parameterization in speech recognizers. Practical applications of speech recognition and dialogue systems bring sometimes a requirement to synthesize or reconstruct the speech from the saved or transmitted MFCCs. Presented paper describes an approach to the construction of a MFCC-based speech production system and discusses various possibilities of its excitation.

1. INTRODUCTION

In recent years thanks to available higher computational power of computers the world leading research teams have begun to use the *mel-frequency cepstral coefficients* (MFCC) [1] as the front-end speech parameterization. This kind of parameterization has been refined mainly for the purpose of speech recognition. Practical applications of speech recognition and dialogue systems bring sometimes a requirement to synthesize (reconstruct) the speech from the saved or transmitted MFCCs.

For the reconstruction of the speech from the mel-frequency cepstrum the model based on straight approximation of log magnitude spectra with mel-scaled frequency by the linear filter was introduced in [2], [3]. Described approach evaluates the output coefficients as the truncated vector of the full cepstra computed from the full log magnitude mel-frequency spectra. However there are differences among algorithms used for the MFCC parameterization in practice, especially in the speech recognition tasks. Main differences arise from how the ‘mel-filtering’ is applied. Usually the mel-filtering is realized in such algorithms as a bank of band-pass filters which performs simultaneously mel-frequency warping and banding to much smaller number of bands. Of course, it brings about the inaccuracy (in comparison with algorithm presented in [3]) which causes significant differences in resulting cepstral coefficients.

To solve this problem it was necessary to espouse the hypothesis that the speech production can be based on the model which models process of hearing (as MFCC parameterization does). It is done with respect to the known relations between characteristics of auditory and speech organs of human beings. The mentioned principle based on the straight approximation by linear filter was successfully probed and accepted for the specific ‘MFCC’ algorithm. The obtained model has proved to be acceptable in practice.

Due to the fact that the information about an excitation is lost by the MFCC evaluating algorithm it is needful to use some additional mechanism to obtain the information required for the production of an excitation. Because of features of the model it is possible to use the method, known as ‘inverse filtering’, to obtain the information (signal) essential for the reconstruction. Such a signal, often called as residuum, can be used straight as an excitation. In such a case the reconstructed speech signal can be considered to be identical as an original natural speech signal. Because the used model introduces some drawbacks caused mainly by its sensitivity in stability to MFCC’s amplitude, the resulting signal can produce time to time certain dropouts. The method minimizing this effect by equivalent modifications of model’s structure was proposed in [4].

2. MFCC EVALUATION

The MFCC parameterization follows common requirements imposed on a speech parameterization for speech recognition purposes. Its main features are aimed above all at:

– to capture an important information presented in a speech signal for recognition purposes
– to handle as little data as necessary and
– to use any quick evaluation algorithm.

Moreover the benefit of MFCCs is also in their perceptually scaled frequency axis. The mel-scale offers higher frequency resolution on the lower frequencies in the same way as a sound is percepted by the human auditory organ. In addition, the MFCCs offer through their cepstral nature abilities to model both poles and zeros.

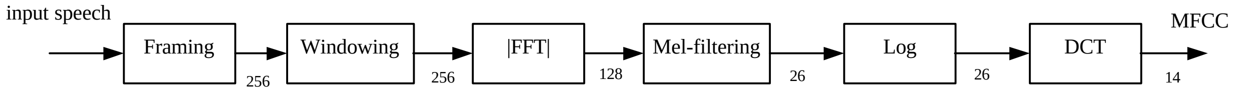

Fig. 1 Block diagram of the MFCC evaluation algorithm (numbers among blocks mean vector lengths).

In Fig. 1 a block diagram of a MFCC evaluation algorithm is shown. The speech signal is first framed to frames the sizes of which are usually chosen as a power of two to fit the FFT algorithm. In the next block samples of a speech signal presented in the frame are weighted by the Hamming window. Then the FFT algorithm is applied to get the magnitude spectrum of the windowed speech data. The next block specified as mel-filtering provides a model of hearing realized by the bank of triangular filters uniformly spaced in the mel scale (Fig. 2a). The mel frequency scale is defined as

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right), \qquad (1)$$

where f is frequency in Hz. Applying the mel filter bank to magnitude spectrum results in high reduction of data amount representing analyzed frame. In our system we use 26 triangular filters. The filters have not equal gain because of their varying frequency widths (in linear scale). As it was verified it does not matter for purposes of speech recognition. But for purposes of speech production it is useful to have each stage of modeling as accurate as possible. So the output of the bank of mel filters is weighted by a correction function (Fig. 2b). In practice each point of that function is obtained as inverse value of a sum of coefficients of individual filters. Then multiplying output of the filter bank by mentioned correction function we would get the mel spectrum as if the filters have unit gain. In our algorithm this correction is not, in fact, applied at this point. This correction can be equally performed in cepstral domain only if it is needed. Following block, which has its origin in classic homomorphic speech processing, performs natural logarithm. The usual role of logarithm in homomorphic processing is to separate convoluted signal's components. In case of speech processing these components usually model the vocal tract and the excitation. In our case the information about excitation is already lost during filtering by the mel filter bank. So the log module acts here only as a smoothing function. The smoothed function (spectrum) is more suitable for cepstral representation. The smoothness causes faster lowering of cepstral components and then possible usage of shorter cepstra. In our system we used 14 cepstral (MFCC) coefficients. The last algorithm stage performed to obtain mel frequency cepstral coefficients is a Discrete Cosine Transform (DCT) which encodes the mel logarithmic magnitude spectrum to the mel frequency cepstral coefficients MFCC.

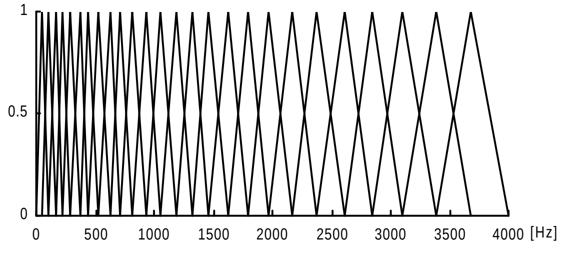

Fig. 2a Frequency responses of the bank of mel filters.

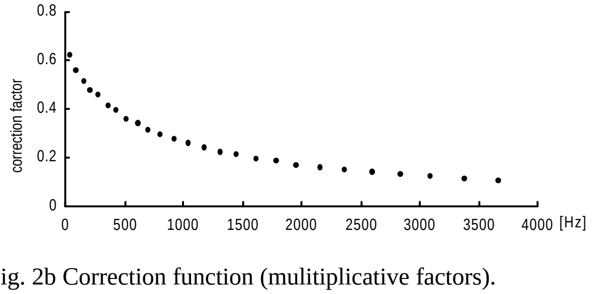

Fig. 2b Correction function (mulitiplicative factors).

3. SPECTRAL ESTIMATION

Before the speech production model will be described it is appropriate to consider a way in which the MFCCs represent the speech spectrum. As shown above the MFCCs are the terms of the cosine expansion of a logarithmic magnitude spectrum (LMS) expressed on the mel-scale. It means that for given number of coefficients the approximation of the mentioned LMS is the best one in the sense of the least square criteria. Fig. 3 shows the resulting approximation of LMS obtained as a sum of cosines. We can see how well (or bad) these coefficients represent speech spectrum and what could be required of the model we attempt to build.

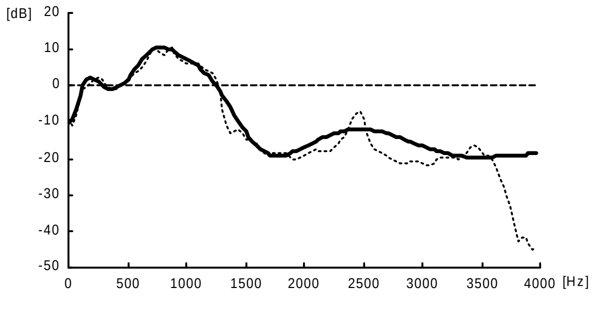

Fig. 3 The original speech spectrum (dotted line) and speech spectrum (solid line) obtained as a sum of cosines.

4. SPEECH PRODUCTION MODEL

The effort to use similar principles for a speech production as in LPC-based system led us to find a model in the form of a digital filter. To reconstruct the speech waveform from the mel-frequency cepstrum we have used the model based on the straight approximation technique of log magnitude spectra for the linear filters [2], [3]. The requirement is to obtain a filter with logarithmic frequency response

$$\log\left|H(\exp(\mathrm{j}\overline{\mathrm{w}}_k))\right| = \sum_{m=0}^{M} \bar{c}_m \cos(\overline{\mathrm{w}}_k m) \,, \qquad (2)$$

where w represents a frequency on the mel scale, the *H* is a transfer function of a desired reconstructing filter, and $\bar{c}_m$ are MFCCs. It is seen that we require to find a digital filter with an exponential frequency response on the mel scale. As it is not possible to build a filter with exactly exponential frequency response, the Padde approximation is used [2] to approximate it by the rational function. Using the z-transform and the 2nd order Padde approximation, the transfer function *H* can be expressed in the form

$$H(z) = \mathrm{e}^{\bar{c}_0} \prod_{m=1}^{M} \mathrm{e}^{\bar{c}_m \bar{z}^{-m}} = \mathrm{e}^{\bar{c}_0} \prod_{m=1}^{M} \frac{1 + a_m \bar{z}^{-m} + b_m \bar{z}^{-2m}}{1 - a_m \bar{z}^{-m} + b_m \bar{z}^{-2m}} \,, \qquad (3)$$

where $\bar{z}$ is a variable of z-transform modified by mel scale, a_m and b_m are the coefficients obtained by Padde approximation. The mel-scaling can be modeled using a quite simple all-pass transform performed by the all-pass filter

$$\frac{Y(z)}{X(z)} = \frac{z^{-1} - \mathrm{a}}{1 - \mathrm{a} z^{-1}} \overset{\text{substitution}}{=} \bar{z}^{-1}, \qquad (4)$$

where constant a shapes the phase frequency response of the filter to the mel scale (1). The relation between a and sampling frequency could be deduced by comparing the phase frequency response of the filter (4) and the mel scale (1).

Unfortunately the substitution (4) used in (3) results in a non-causal and non-realizable filter. To solve this problem we applied the technique published in [3]. First we rewrite the equation (3) to the vector form

$$H(z) = \exp(\bar{\boldsymbol{z}}^{\mathrm{T}} \bar{\boldsymbol{c}}) \,. \qquad (5)$$

Then applying the all-pass transform expressed by the matrix ***T***

$$\boldsymbol{T} = \begin{bmatrix} 1 & \mathrm{a} & 0 & \Lambda & 0 \\ 0 & 1 & \mathrm{a} & \mathrm{O} & 0 \\ 0 & 0 & 1 & \mathrm{O} & 0 \\ \mathrm{M} & \mathrm{O} & \mathrm{O} & \mathrm{O} & \mathrm{a} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad (6)$$

to the vectors $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{c}}$

$$\bar{\boldsymbol{z}}^{\mathrm{T}} \bar{\boldsymbol{c}} = \bar{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{T}\boldsymbol{T}^{-1} \bar{\boldsymbol{c}} \qquad (7)$$

we obtain the transfer function in the form

$$H(z) = \exp \sum_{m=0}^{M} d_m \ddot{O}_m \,, \qquad (8)$$

where

$$\ddot{\boldsymbol{O}}(z) = \bar{\boldsymbol{z}}^{\mathrm{T}} \boldsymbol{T} \quad \text{and} \quad \boldsymbol{d} = \boldsymbol{T}^{-1} \bar{\boldsymbol{c}}. \qquad (9)$$

Applying the Padde technique to (8) to approximate the exponential we obtain the filter, which is realizable because of

$$_m(z) = \left\langle \begin{array}{ll} 1 & m = 0 \\ z^{-1}\left[\dfrac{1-\mathrm{a}^2}{1-az^{-1}}\right]\bar{z}^{-(m-1)} & m = 1,\dots, M\,. \end{array} \right. \qquad (10)$$

Now we have obtained the filter, which models through its magnitude frequency response the respective frame of speech. Fig. 4 offers example of its magnitude frequency response to compare it with the speech spectrum (dotted line) obtained as a sum of cosines (solid line in Fig. 3).

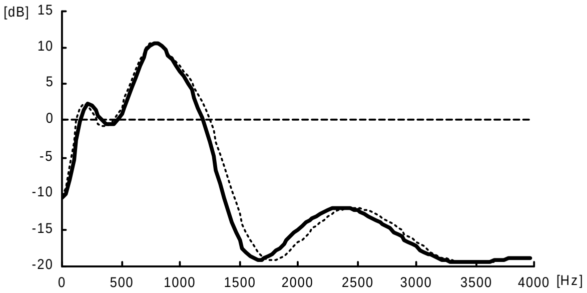

Fig. 4 The magnitude frequency response (solid line) compared with the speech spectrum (dotted line) obtained as a sum of cosines (as in Fig. 3).

5. EXCITATION

Before the problem of a model excitation will be discussed let‘s consider that the MFCC-based model has especially been developed for purposes to model human hearing. The hypothesis about a possibility to produce a real speech by means of a model of hearing results from a partial similarity of these two processes. Note that if the magnitude frequency response of a model of hearing sufficiently copied the corresponding response of a real speech production system then we could get the perfect speech signal produced by the model using the excitation with an exactly flat spectrum. Unfortunately, as the frequency attributes of the hearing and speech production process are not quite identical, we cannot

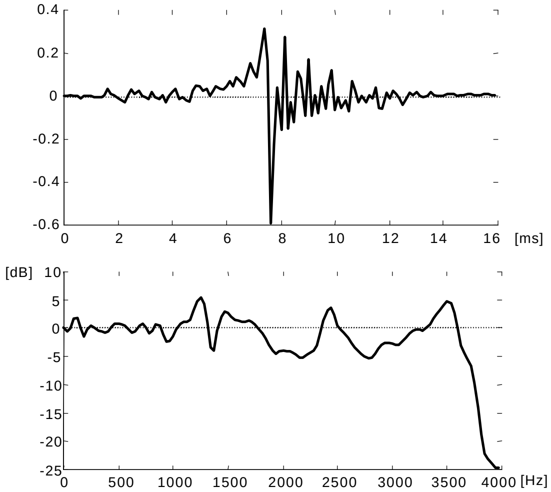

Fig. 5 The residual signal and its magnitude spectrum

submit any straight hypothesis about an excitation of a model of hearing to produce a real speech signal.

The question how to excite mentioned model can be answered analyzing the residual signal obtained by the technique of an inverse filtering. This approach is possible due to the fact that the model can be considered to be stable with a minimum phase. Such features of the model need not be satisfied automatically but depend on absolute values of cepstral coefficients. To ensure the stability we built the models as the cascade filters of a lower degree [4]. Figure 5 shows the resulting residual signal and its magnitude spectrum for the same frame of a speech that was processed and depicted in the preceding figures. In spite of the local peaks the residual magnitude spectrum can be considered to be sufficiently flat. This result was confirmed by many further experiments. So the hypothesis about using the model of hearing to the speech production could be admitted to be valid:

– the characteristics obtained from MFCCs representing the model of hearing are close enough to the characteristics of the real speech production (at least in the magnitude frequency domain) and
– the flatness of residual spectrum is sufficient at least for lower frequencies.

The speech reconstruction process can be performed using various excitation signals. For the speech of a high quality the full residual signal may be applied. Of course a usability of such excitation will be somewhat restricted in real applications due to the need to transmit and/or to store large number of excitation data (residual signal). If no residual signal for the excitation of the model is in disposal (only MFCCs are stored in a memory and/or there is a limited width of a transmission channel), a speech signal has to be reconstructed only from MFCC vectors including some simple excitation deduced not from real residual signal. Note that even the very simple excitation like a pulse train with constant period allows “to replay the MFCCs” as an intelligible resulting speech sound. It is evident that between two mentioned performances of an excitation (full residual signal or pulse train with a constant period) many other “middle-quality” speech reconstruction approaches could be found by providing some additional information to the MFCCs. For such purpose the Hilbert transform of a unit pulse may be used, which offers wide band spectral excitation. An improvement in a speech production process could also be achieved extracting a real pitch period from an analyzed speech and joining it as an extra feature to the MFCCs. We also performed several experiments with a DFT coding of a residual signal, an additional cepstral analysis of residuum and multi-pulse excitation. These experiments haven’t been finished yet and will have to be completed with subjective tests to infer serious conclusions. Generally we can say that techniques for improving an excitation to reach more natural speech output are nearly same as in LPC-based speech production systems.

6. CONCLUSION

The speech production (reconstruction) system based on the MFCCs is supposed to be usually used in those applications which impose limitations mostly on a width of a transmission channel and/or a memory size (there are/have to be stored or transmitted only MFCC parameters).

In our latest research we pay an attention to the development of a pitch-synchronous TTS synthesis system based on the MFCCs. We decided to use full residual excitation signal to excite a concatenation-based synthesizer using triphones as the basic speech units.

7. ACKWNOLEDGEMENT

This research is supported by the Ministry of Education of the Czech Republic - grant No. VS 97159, and by the Grant Agency of the Czech Republic - project No. 102/96/K087.

8. REFERENCES

[1] DAVIS S., MERMELSTEIN P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP, ASSP-28, 1980, 357-366.

[2] IMAI S., KITAMURA T., TAKEYA H.: A direct approximation technique of log magnitude response for digital filters. IEEE Trans. on ASSP, ASSP-25, 1977, pp. 127-133.

[3] IMAI S.: Cepstral analysis synthesis on the mel frequency scale. -In: Proceedings ICASSP-83, 1983, pp. 93-96.

[4] PØIBIL J.: The use of the cepstral model for speech synthesis. Thesis. Czech tech. university. Prague, 1997. (in Czech)