PITCH SYNCHRONOUS RESIDUAL EXCITED SPEECH RECONSTRUCTION ON THE MFCC

Zbyněk Tychtl and Josef Psutka
University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
e-mail: tychtl@kky.zcu.cz, psutka@kky.zcu.cz

ABSTRACT

Practical applications of speech recognition and dialogue systems bring sometimes a requirement to synthesize or reconstruct the speech from the saved or transmitted mel-frequency cepstral coefficients (MFCCs). Presented paper describes several approaches to the speech reconstruction based on the MFCC parameterization. Approaches differ mainly in their various possible excitations. Let us mention that for designing a MFCC reconstruction module we applied some principles usually used in speech recognition and synthesis process. We suppose the speech reconstructor together with speech synthesizer and recognizer to be a part of a speech dialogue system developed in our department.

1 INTRODUCTION

In our speech recognition systems we use FFT based mel-frequency cepstral coefficients (MFCC) [1] as the front-end speech parameterization. The efficiency demands bring us to the need to be able also to reconstruct the speech from that parameterization. In [2], [3] the model based on straight approximation of log magnitude spectra with mel-scaled frequency by the linear filter was introduced. This model can be used for the reconstruction of the speech signal from the mel-frequency cepstrum. Described approach evaluates the output cepstral 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, this technique brings about the inaccuracy (in comparison with algorithm presented in [3]), which causes significant differences in resulting cepstral coefficients.

The principle mentioned in [3] that is based on the straight approximation by linear filter was successfully probed and accepted for the specific ‘MFCC’ algorithm. The obtained model [5], [6] 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 necessary 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 excitation 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. Since the used model introduces some drawbacks caused mainly by its sensitivity of stability to MFCC’s amplitude, the resulting signal can produce from time to time certain dropouts. The method minimizing this effect by equivalent modifications of model’s structure was proposed in [4].

Firstly the simplest version of speech reconstruction system is presented. Here it is not supposed any additional information to be available besides the sequence of MFCC vectors representing the speech signal to be reconstructed. As the excitation we use simple pitch train generator for the voiced sounds and noise generator for the unvoiced sounds. The voiced/unvoiced decision is taken on the MFCC vector only. In this case it is impossible to sign described technique as a pitch synchronous approach because there is nothing to be synchronous with.

Significantly better results can be achieved when we are able to obtain voiced/unvoiced flag or (even and) F0 or even the pitch marks from the MFCC parameterizator. In this case when we know F0 the pitch marks can be calculated to get pitch synchronous reconstruction.

If we go a step further we aim to design not only pitch synchronous but also residually excited reconstruction to get spectrally rich and zero-modeling reconstructed signal. For that purpose we have to consider the questions of residual excitation preparation, its storing and appropriate matching with incoming MFCC vectors. The proposed methods will be discussed in the paragraph 3.

The same approach to speech signal generation we are using for purposes of speech synthesis in our TTS system [7], [9].

2 MFCC BASED PRODUCTION MODEL

The MFCC parameterization has been refined to follow common requirements imposed on a speech parameterization for speech recognition purposes. Its main features are aimed above all at:

- to capture an important information present 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.

In the MFCC calculation 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 stage the samples of a speech signal presented in the frame are weighted using the Hamming window. Then the FFT algorithm is applied to get the magnitude spectrum of the windowed speech data. The mel-filtering provides a model of hearing realized by the bank of triangular filters uniformly spaced in the mel scale. The mel frequency scale is defined as

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

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. 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 get the mel spectrum as if the filters have unit gain. In our algorithm this correction is not, in fact, applied at this point. It can be adequately performed in cepstral domain only if it is needed. Following stage, which has its origin in classic homomorphic speech processing, performs the natural logarithm. The usual role of the logarithm in homomorphic processing is to separate convoluted signal components. In case of speech processing these components usually model both 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 logarithm acts here only as a smoothing function. The smoothed function 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 use 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. It means that the MFCCs are the terms of the cosine expansion of a logarithmic magnitude spectrum expressed on the mel-scale.

To find out the reconstruction model in a form of digital filter 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 $H(z)$ with a logarithmic frequency response

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

where $\bar{\omega}$ represents a frequency on the mel scale, the H is a transfer function of a desired reconstructing filter, and $\bar{c}_m$ are MFCCs. The Padde approximation is used [2] to approximate exponential by the rational function. The mel-scaling can be modeled using a quite simple all-pass transform performed by the 1^{st} order all-pass filter.

Now we have the filter, which models through its magnitude frequency response the respective frame of speech. 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].

The MFCC parameterization has especially been developed for purposes to model human hearing. Unfortunately, as the frequency attributes of the hearing and speech production process are not quite identical, we cannot 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 by analyzing the residual signal obtained by the technique of an inverse filtering. In spite of the local peaks the residual magnitude spectrum can be considered to be sufficiently flat for most of cases. 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.

3 SPEECH RECONSTRUCTION

The main motivation for this research was the need to 'replay' pronunciations encoded into the mel-frequency cepstral coefficients obtained by the mentioned parameterization algorithm. It is obvious that in such systems we will rarely have latitude to change or extend the used parameterization process to extract any additional information essential for as good reconstruction as possible. Basically it means that we are restricted to use only the MFCCs. Using a 'filter-like' model to the speech production the excitation signal need to be generated. We have performed some experiments with the goal to find out any 'universal' excitation independent on voice feature of the speech segment. The results showed that we would get better quality of reconstructed speech if we had also the voiced/unvoiced information. It is straightforward that it would be very useful to be able to make this 'voiced/unvoiced' decision directly on the basis of mere MFCCs vector.

We have decided to look for any criterion or rule helping us to make such decision. We have found out very simple criterion, which performs mostly even better than classic autocorrelation methods. The basic idea was that it should be possible to make this decision according to the 'shape' of the spectral envelope and the knowledge how the MFCC vector represents this envelope. It was found that it is sufficient to sum the amplitudes of several first MFCC coefficients (without $\bar{c}_0$) and compare this sum with the preset threshold. It must be said that the number of coefficients to be summed as well as threshold's value depend strongly on the attributes of actual MFCC parameterization. In our case when we use 16kHz sampling frequency, 26 mel filters and 14 cepstral coefficients (including $\bar{c}_0$), it was experimentally verified that it is suitable to sum first 6 coefficients ($\bar{c}_1 ... \bar{c}_6$) and to preset the threshold to 1.5

$$\sum_{j=1}^{6} |\bar{c}_j| > 1.5 \Rightarrow \textit{voiced}. \qquad (3)$$

If the sum is greater then threshold the segment is marked as voiced. The key to the success of this simple method is the 'correction' mentioned above eliminating non equal gains of mel filters. This straightens the tilt of spectrum represented by the MFCC vector and causes that the spectra of voiced and unvoiced segments will be more distinguishable throw their MFCCs. Let it be noticed that mentioned correction does not mean the necessity of a change of the parameterization. It can equally be performed in the cepstral domain as the simple vector substraction. One must also keep on mind the influence of an optionally used liftering. If using liftering in parameterization so the MFCCs must either be delifteréd or the 'delifteríng' must be incoporated into the sum (3). The criterion (3) works well when using the amplitude spectra. Threshold's value is determined experimentally and depends on all attributes of parameterization.

Now we have got the method for making the voiced/unvoiced decision. It allows us to use different excitations in these two cases, as it is common in other model-based approaches to speech production. The simplest excitation we use can be pulse train for voiced segments and white noise for unvoiced ones. An improvement could be experienced using some composite excitations. We use the impulse response of the Hilbert transformer to excite the voiced segments. This simple signal offers several features useful for pitch synchronous speech production. It has exactly flat spectrum in the segment. The energy is not concentrated in one pulse but is spread over whole segment and the main pitch is located in the center of the segment. This is useful for PSOLA algorithm.

To perform the pitch synchronous reconstruction we employ the adopted PSOLA algorithm. Slight difference is in fact that it is not applied to the speech signal but to the excitation. By this way accordingly to incoming MFCCs the excitation is built from voiced and unvoiced excitation segments. These segments have double pitch period length and are Hanning windowed. Currently the algorithm uses a pitch period preset by the user. Let's notice that speech parameterization (MFCC) is not supposed to be obtained as pitch synchronous, since it is rare in speech recognition systems. It means that incoming MFCC vectors are asynchronous with pitch period. To get quality of reconstruction the pitch synchronous process has to have the priority over the exact timing of the train of parameters. Another improvement is possible by interpolation of neighboring MFCCs. There can be used quite simple interpolation method thanks to MFCCs' cepstral nature. The interpolation is especially beneficial to overcome dropouts arising due to parameters switching. Note that even the very simple excitations like a pulse train with constant period and a noise allow 'replaying' the MFCCs as an intelligible speech sound.

To reach higher quality speech reconstruction 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) that is of the same amount as the original speech signal. The terms 'residuum' and 'residual signal' designate here the excitation signal that can be obtained by the method known as inverse filtering. Thanks to the production model's features it is possible to build the inverse model. Exciting the inverse model by the original speech signal we get mentioned residual signal. Using this 'full' residuum to excite the production model we obtain the reconstructed speech equal to the original speech signal. This is true when the model is stable with a minimum phase.

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 excitation deduced from real residual signal may be used. Proposed paper does not cover these middle-quality approaches.

In presented new quality reconstruction system we suppose no additional information besides the sequence of MFCCs and the F0 to be available. To achieve the high quality speech signal we insist on using a pitch synchronous residual excitation. It means that we have to have some database of prepared residual excitation units and mechanism for choosing them appropriately according to the incoming MFCC vectors. On chosen residual excitation units the PSOLA algorithm is applied to obtain continuos excitation signal. Pitch mark generator that uses F0 to synchronize itself controls the overlapping in the PSOLA algorithm. This composed residual signal is then used as excitation of a production filter which structure and coefficients are set according to MFCC coefficients.

The principles of a residual excitation unit database preparation are based on speech recognition methods. The database was derived from the acoustic unit inventory that was automatically prepared using hidden Markov modeling [8]. Hidden Markov Models (HMMs) are used to model triphones. The states of triphone HMMs are automatically clustered down using binary decision trees. The clustered states are then used to automatically segment the speech corpus and to create a speech segment database. Each speech segment in this database is tracked for pitch marks (in case of voiced segments) and then one frame of double pitch period is picked and Hanning windowed. Also the MFCC vector is calculated for this frame. Then the residual signal is obtained by its inverse filtering and finally this new frame of residual signal with corresponding MFCC vector is stored to the database of residual excitation units.

Another way of choosing the vectors and segments is based on a statistical approach. For each cluster one representative MFCC vector is computed as the average of all participating vectors. For each segment in the cluster the residual signal is obtained by inverse filtering. All these residual segments are then pitch synchronously cut up. Their spectra are then determined and an average spectrum is provided. In fact we get an average estimation of the amplitude residual spectrum. The phase of the nearest segment’s spectrum is then embedded as the phase in this average spectrum. Then using the inverse DFT we obtain the time representation of the excitation segment of respective cluster. This averaging method is legitimate since residuals are as well as speech signals stationary. Then for each cluster a representative MFCC vector and a respective excitation segment are stored in a database. In the process of reconstruction each incoming MFCC vector is compared with those stored in database, the nearest one is found and respective excitation segment is used. To choose the nearest vector in the database the Euclidean distance is used.

During a presentation listening tests will be performed to compare proposed methods.

4 ACKWNOLEDGEMENT

This research is supported by the Ministry of Education of the Czech Republic - projects no. VS 97159 and no. MSM 23520004, and by the Grant Agency of the Czech Republic - project no. 102/96/K087.

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 of the ICASSP’83, 1983, pp. 93-96.

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

[5] TYCHTL Z.: Model for Speech Production Based on the Mel Cepstral Coefficients and Its Applications. Rigorous report. (in Czech) Plzeň, 1998.

[6] TYCHTL Z., PSUTKA J.: Speech Production Based on the Mel-Frequency Cepstral Coefficients. -In: Proceedings of Eurospeech’99, pp. 2335-2338, Budapest 1999.

[7] MATOUŠEK J.: Speech Synthesis Using HMM-Based Acoustic Unit Inventory. -In: Proceedings of Eurospeech’99, pp. 2323-2326, Budapest 1999.

[8] MATOUŠEK J., PSUTKA J., TYCHTL Z.: Statistical Approach to the Automatic Synthesis of Czech Speech. -In: Proceedings of the 2nd International Workshop on Text, Speech and Dialogue TSD’99. Springer Verlag, Berlin 1999, pp. 376-379.

[9] MATOUŠEK J., PSUTKA J., MÜLLER L.: Text-To-Speech Synthesis Using HMM-Based Triphones. -In: Proceedings of the 10th Annual International Conference on Signal Processing Applications and Technology ICSPAT’99. Orlando, U.S.A., 1999.