Cepstrum-based estimation of resonance frequencies (formants) in high-pitch singing signals

C. Zarras[1], K. Pastiadis[2], G. Papadelis[3], G. Papanikolaou[4]

[1] *Aristotle University of Thessaloniki, 54124 Thessaloniki, E-Mail: chzarras@auth.gr*
[2] *Aristotle University of Thessaloniki, Dept. of Music Studies, 540 06 Thessaloniki, E-Mail: pastiadi@mus.auth.gr*
[3] *Aristotle University of Thessaloniki, Dept. of Music Studies, 540 06 Thessaloniki, E-Mail: papadeli@mus.auth.gr*
[4] *Aristotle University of Thessaloniki, 54124 Thessaloniki, E-Mail: pap@eng.auth.gr*

Abstract

The estimation of the vocal tract resonance frequencies from acoustic voice signals has been widely employed and various methods have been proposed. Among them, a number of cepstrum based techniques have been implemented to disentangle the voice's spectral envelope from the harmonic components. Noticebly less research has been conducted for voices with higher fundamental frequency, as in singing (e.g. soprano voices). In such cases, the estimation of the spectral envelope is affected by the presence of cepstral rahmonics, which are interleaved with spectral envelope estimation. In this paper, some new techniques based on cancellation of rahmonics, rather than hard liftering, are proposed and examined for their effectiveness in maintaining the spectral envelope information. Both straightforward implementations and iterative procedures are considered and simulation results for various configurations of f_0 and formant frequencies are presented. These preliminary examinations allow the evaluation of effects of various acoustical and signal processing factors on estimation accuracy and assess the feasibility of the proposed approaches for use with high fundamental frequency signals, such as singing, and in other similar fields of interest in musical acoustics.

Introduction

Formants are defined as the resonance frequencies of the stomatopharyngeal filter, or in other words the local maxima of the filter's transfer function. Formant frequency estimation is a common task in speech processing and is of great interest due to its application to various fields such as speech encoding, speech and speaker recognition.

For harmonic speech signals with low fundamental frequency, harmonics are close enough to each other and formants can be obtained through spectral envelope's peaks. In contrary, when f0 exceeds a high value, harmonic distance increases and estimating an accurate representation of the stomatopharyngeal filter's transfer function, becomes more complicated. In such cases, spectral maxima and formants are not necessarily coinciding. This problem is apparent in women and children's singing signals and even more in soprano voices whose fundamental frequencies exceed 1 kHz.

A number of methods have been proposed for the estimation of formants. Among them, the LPC method [1] which is based on linear prediction of speech, as well as some cepstrum-based techniques [2] which are based on the signal's real cepstrum. However, for both types of methods, there are disadvantages which become apparent as fundamental frequency increases[3]. Formants estimated with LPC tend to follow the spectral peaks, i.e. the harmonics, ignoring the true vocal tract resonances. From the other hand, in high pitch signals, cepstral rahmonics coexist at the lower part of the cepstrum with spectral envelope information making the disentanglement rather difficult. Other methods based on 2D time- frequency representations [4] or the true envelope approach [5] have also been proposed.

In this preliminary study, a new approach in cepstral liftering is presented and tested with synthetic voice signals. The aim is to minimize the cepstrum drawbacks when used with high pitched voice signals. A new Voice Generation Application was developed for the tests based on the LF model [6]. The test results and prospects of further investigation are discussed.

Methods and Results

The limited usage of cepstrum-based algorithms in high pitch singing signals is due to the characteristics of the cepstral representation. The main principle of these methods is that the spectral envelope information is mainly located at low order cepstral coefficients (quefrencies), while harmonics information goes up to higher quefrencies. This makes their disentanglement easy. As the fundamental frequency of the tested signal increases, the distance between the first rahmonic and the start of the cepstrum and between the rahmonics each other decreases. Given that the lower quefrencies are necessary for the extraction of spectral envelope information, setting cepstral coefficients at rahmonic positions to zero is needed. Consequently, the presence of rahmonics in that low area and their cancellation distorts the envelope.

Our approach relies on the fact that lower rahmonics are more affected by spectral envelope information than higher ones. After all, this is the main reason why the lowest part of the cepstrum is used for the calculation of the smoothed spectral envelope. Therefore, one of the latest rahmonics, which are considered "clearer" i.e. their value is less affected by the filter's transfer function can be subtracted from the first rahmonics, in order to minimize the harmonics information in the final smoothed spectral envelope representation.

The algorithm requires a decent estimation of pitch frequency as a first step, since rahmonics have to be located.

The procedure deployed for the pitch estimation, calculates the autocorrelation function of the frame's cepstrum. The f_0 value is then estimated using the following equation:

$$f_0 = \frac{f_s}{D} \quad \text{[Hz]} \qquad (1)$$

where f_s is the sampling frequency, and D is the distance between two successive maxima of the autocorrelation function.

The next step is called the rahmonic subtraction. The purpose of this step is to eliminate the harmonic information from the cepstrum preserving only the spectral envelope information. Following the rationale described above, a higher order rahmonic (i.e. one that is located above the quefrency limit that defines the part of the cepstrum that will be used for the smoothed envelope estimation), is selected and subtracted from all rahmonics that are within the lower cepstral part of interest. Finally, the smoothed spectral envelope is estimated after setting to zero all quefrencies outside this part. The algorithm is summarized in the following diagram:

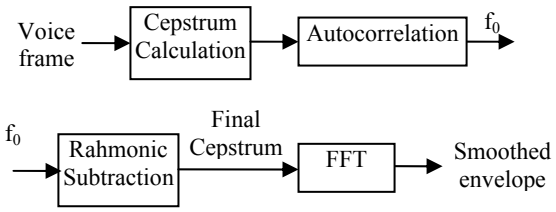

Figure 1: Block diagram of the deployed algorithm

In order to examine the effectiveness of the algorithm in this preliminary phase, synthetic voice signals were used. A new voice synthesis application was developed, based on the linear separable model. The glottal signal is generated using the LF model, while the stomatopharyngeal filter is implemented by up to six bipolar filters in a row, with to conjunctive poles each. The number of bipolar filters determines the number of formants in the final signal. The application allows more than forty parameters to be determined, such as glottal pulses shape, formant frequencies and bandwidths, noise types and lengths, jitter, shimmer etc. Only vowel-like synthetic voice signals can be generated by the application. The true and estimated formant frequencies for four different fundamental frequencies are as follows. The duration of the tested signals was 50ms each:

Table 1: True and Estimated Formant Frequencies (Hz)

f0	F1	F1 estimated	F2	F2 estimated
400	600	620,8	1500	1522,1
500	700	700,95	1800	1842,5
700	900	919,5	1900	1879,5
800	950	961,3	2000	1982,7

Figure 2 shows an example of the spectral preservation for a 700Hz signal. The red line is the true spectral envelope, while the blue one is the spectral envelope estimation.

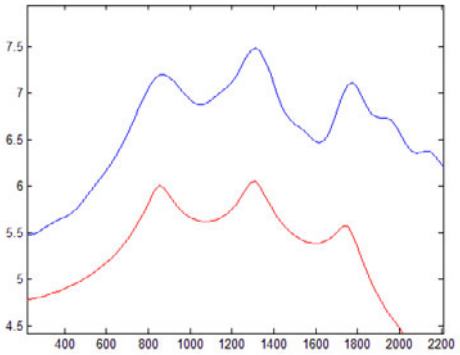

Figure 1: Block diagram of the proposed algorithm

As we can see from these examples, the spectral envelope estimation is quite accurate even when high pitch voice signals are used. The accuracy of the envelope is better for the two lower formants.

Conclusions

The aim of this paper is to deploy a cepstrum-based method using rahmonic subtraction for the estimation of formant frequencies in high-pitch voice signals. The method was tested using only synthetic voice signals. Although cepstral techniques are, generally, not suitable for high pitch voice signals, the results were promising since quite good estimations of the two lower formant frequencies at least were made. However, further investigations using real voices should be made before extracting safe conclusions.

Bibliography

[1] Markel, J.D. Gray, A.H.: Linear Prediction of Speech, Springer Verlag, 1976.

[2] Oppenheim, A.V. Schafer, R.W.: Digital Signal Processing, NJ: Prentice Hall, 1975.

[3] Kammoun, M.A. Gargouri, D. Frikha, M. Ben Hamida, A.: Cepstral Method Evaluation in Speech Formant Frequencies Estimation, 2004 IEEE International Conference on Industrial Technology, vol 3, p 1612-1616, 2004.

[4] Wang, T.T. Quatieri, T.F.: Exploiting Temporal Change of Pitch in Formant Estimation, ICASSP 2008, p 3929 – 3932, 2008.

[5] Robel, A. Rodet, X.: Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation, Proceedings of the 8th International Conference on Digital Audio Effects, p 30-35, 2005.

[6] Zarras, X.: Tools Development and Evaluation of Methods for Formant Frequency Estimation (MSc Thesis), Aristotle University of Thessaloniki, 2009.