AN ALGORITHM FOR PITCH DETERMINATION OF SPEECH BASED ON AN AUDITORY SPECTRAL TRANSFORMATION

M. Beham and W. Datscheweit

Lehrstuhl für Datenverarbeitung
Technical University of Munich

ABSTRACT - An algorithm for pitch calculation is described which is based on an auditory spectral transformation and the theory of virtual pitch perception. The principle of harmonic coincidence is used to estimate pitch frequency.

INTRODUCTION

This contribution deals with a system for calculation of pitch frequency based on the theory of virtual pitch. Pitch extraction is therefore also possible for band-limited signals which do not include the fundamental frequency itself (e.g. telephone signal). The system is divided into three stages. In a first part spectral transformation of the time signal is carried out, which takes into account basic properties of the auditory system. The second stage extracts the relevant tonal components for the perception of virtual pitch, the part tones, and in the last stage pitch frequency is calculated using the principle of harmonic coincidence. The presented system is based on a publication by Terhardt (1982), but in contrast to the system described there it offers an auditory-based spectral analysis and a faster algorithm for pitch calculation.

SPECTRAL TRANSFORMATION

In the first stage the perceptually based auditory representation of the discrete time varying signal s(i) is computed using a system (Beham 1990) as displayed in fig. 1. For every frequency point Ω_n the calculation of the complex short-time spectrum is realized by a digital bandpass BP_n (mid frequency Ω_n). The equivalent analog impulse response of the bandpass BP_n, which ensues directly from the used time window $w_n(t)$ of short-time analysis, is $w_n(t)*\exp(-j\Omega_n t)$. Because only the magnitude of the spectrum is auditorily relevant every complex bandpass BP_n is followed by a unit to calculate the square of the magnitude. The output of this stage, the short-time power spectrum, is lowpass filtered for temporal averaging and downsampled to a rate of one spectrum every 5 ms. Finally the logarithm of every spectral value is calculated. The output of this system is an array with N elements for each downsampled time point k. Each element represents the logarithmic short-time power spectrum at a certain frequency Ω_n (n=1,..,N) in dB.

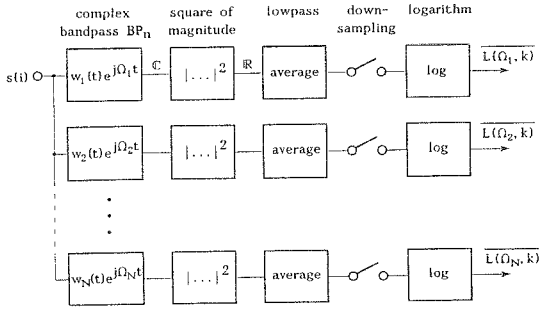

Figure 1. Overview of the spectral transformation stage

Two main effects of the auditory system, the variable frequency resolution power and the nonlinear auditorily based distribution of the points Ω_n along the frequency axis (Zwicker 1982) have to be considered in the spectral transformation. In the presented system the mid frequencies Ω_n of each bandpass BP_n are therefore equally spaced on the auditorily relevant bark scale. The N (N=128) frequency points Ω_n have a constant distance of 0.15 bark along a bark scale between 20 Hz and 6.0 kHz. On the other hand the variable frequency resolution can be imitated by varying the bandwidth of each bandpass BP_n dependent on its mid frequency Ω_n. According to the model of the basilar membrane (Flanagan 1972) the time window $w_n(t)$ for each bandpass was selected as the impulse response of a third order lowpass filter, having the form:

$$w_n(t) = t^2 * \exp(-a_n t)$$

The parameter a_n in the presented time window allows the bandwidth of each bandpass BP_n to be adjusted to the frequency resolution of the auditory system by varying the effective length of each time window $w_n(t)$ as a function of the frequency point Ω_n. With respect to the critical bandwidth of the auditory system (as described by the bark scale, Zwicker 1982) the value of each a_n was estimated to achieve a bandwidth of 0.5 bark.

The other parameters of the spectral transformation were selected as follows:

- sampling frequency of time signal = 16 kHz
- downsampling factor = 80
- quantisation of time signal = 12 bit
- 3dB-frequency of lowpass = 50 Hz

EXTRACTION OF TONAL COMPONENTS

In fig. 2a a spectrum from the center of the vowel /a:/ derived from the previously described analysis is plotted along the bark scale. The harmonic tonal components of the speech signal, the part tones, are represented by local maxima in the spectrum up to about 1.5 kHz.

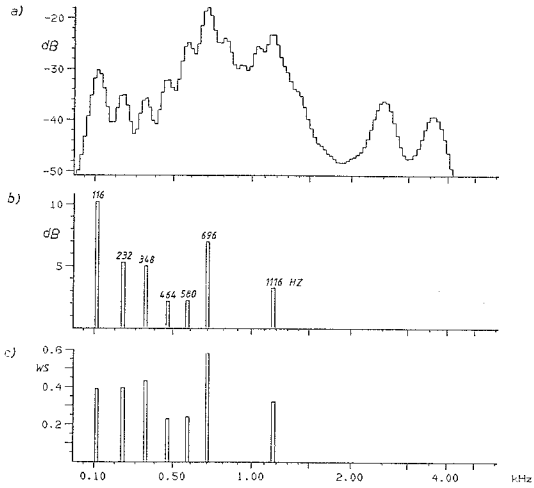

Figure 2a-c. Short time spectrum (a), level excess LE_i (b) and weighted level excess W_i (c) of relevant part tones from the center of the vowel /a:/

The auditorily relevant part tones are detected as local maxima in the spectrum using a certain threshold criterion. The strength of a maximum has to achieve a level excess higher than 2 dB against the left and right neighbouring minima, to be extracted as a tonal component. The level excess of a maximum is approximated by the level difference between the maximum and the one of the two neighbouring minima having a higher level. The accuracy of every part tone's location on the frequency axis f_i is improved using a parabolic interpolation. Fig. 2b shows the extracted maxima and their interpolated frequencies from the spectrum in fig. 2a. The extent to which a tonal component contributes to the entire tonal percept depends on its level excess and its frequency. Due to this reason each extracted maximum from fig. 2b is weighted using the principle of spectral prominence. According to Terhardt (1982) the following slightly modified formula is used to calculate the spectral weight of each tonal component (with LE_i in dB and f_i in kHz):

$$W_i = [1-\exp(-LE_i/8)]*[1+(f_i/0.7+0.7/f_i)^2]^{-0.5}$$

The result is a spectral weight W_i for each part tone f_i in a range between 0 and 1. Fig. 2c shows the tonal components with their spectral weights which are the input for the next stage, the calculation of pitch.

CALCULATION OF PITCH

The presented algorithm estimates the harmonic coincidence of the part tones by compressing the frequency axis with an integer multiple m (see fig.3). The principle behind this is that the "mth" harmonic coincides with the fundamental frequency if its divided by (m+1). The sum of weights and the number of coinciding tonal components is a measure for the significance of the pitch at this point of the frequency axis.

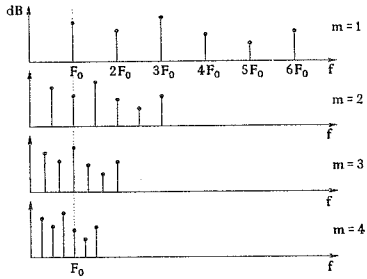

Figure 3. Schematic diagram of harmonic coincidence principle

In practice small errors occur during the interpolation of part tone frequencies f_i. In order to tolerate a relative error δ for calculation of the subharmonic f_i/m the frequency axis is divided into intervals with the length $2\delta F_j$ which are centered around the possible pitch frequency values F_j. Lining up the intervals on the frequency axis leads to a logarithmic scale for pitch values F_j. A part tone f_i is called coincident with the possible pitch frequency F_j if the subharmonic f_i/m (m=1,..,M) fits into the interval around F_j as the following formula shows:

$$F_j - \delta F_j \leq f_i/m \leq F_j + \delta F_j$$

All the spectral weights W_i of coincident part tones f_i and the number of coinciding part tones for each interval j are added, resulting in the sum of weights SW_j and the number of coinciding part tones N_j for each possible pitch frequency F_j. One problem arises with the distinct coincidence intervals. If the unknown pitch frequency F_0 comes close to the boundary between two intervals, some part tones may

coincide with the left and some with the right interval. For that reason two neighbouring intervals are summarized by adding the sum of weights SW_{j+1} or SW_{j-1} (depends on which one is bigger) to those of interval j. Afterwards a maximum search across all intervals j for the highest sum of weights SW_j is carried out and the according F_j is selected as the unknown pitch frequency. The relative harmonic coincidence SW_j/SW_{total} where SW_{total} equals the sum of all weighted part tones W_i is a measure for the significance of the pitch frequency F_j. Its possible range is between 0 and 1, where a value of 1 indicates that all selected part tones are harmonically coincident, i.e. the significance of the calculated pitch is 100%.

EXPERIMENTAL RESULTS

In the presented experiments the possible range for pitch values was limited from 70 to 350 Hz and δ was set to a value of 0.01. Thus the frequency scale between 70 and 350 Hz can be divided into 81 different intervals and therefore also 81 different pitch values are possible. A maximum number of M=10 subharmonics proved to be sufficient. Fig. 4 shows the results of pitch calculation with the weighted part tones from fig. 2c. The relative harmonic coincidence is diplayed for each possible F_j whose number of coinciding part tones is higher than 2. In this example the resulting pitch is F_0=115Hz with a harmonic coincidence of 1.0.

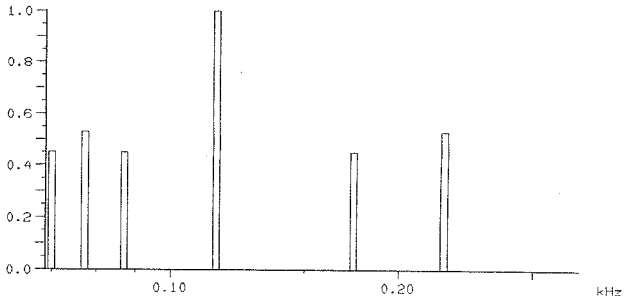

Figure 4. Calculated relative harmonic coincidence from the data in fig. 2

Further experimental results show that if the relative error δ equals 0.01 all detected part tones of vowels are recognized as subharmonics of one fundamental frequency, i.e. the significance of this pitch is always 100%. In the case of voiceless consonants sometimes part tones which are randomly harmonic are detected but a relative coincidence of 0.5 is never exceeded. Difficulties arise with the pitch estimation of voiced consonants. Because the energy of the speech signal is often very low the extraction of part tones which achieve a level excess of 2 dB fails (except the detection of the part tone representing the fundamental frequency itself). Therefore pitch cannot be calculated exactly in these cases. A temporal tracking of pitch frequency using a steadiness criterion may improve this disadvantage. But in spite of these facts the presented algorithm offers a fast and secure possibility for calculation of pitch based on the theory of virtual pitch.

ACKNOWLEDGEMENT

This work was carried out in the SFB 204 "Gehör", which is supported by the Deutsche Forschungsgemeinschaft (DFG).

REFERENCES

Beham, M. (1990) *Gehörgerechte Analyse von Sprachsignalen - rekursive Berechnung, Merkmalsextraktion und graphische Darstellung*, Diplomarbeit am Lehrstuhl für Datenverarbeitung, Technical University of Munich.

Flanagan, J.L. (1972) *Speech analysis synthesis and perception*, (Springer Verlag: Berlin)

Terhardt, E.,Stoll, G. & Seewann, M. (1982) *Algorithm for extraction of pitch and pitch salience from complex tonal signal*, J. Acoust. Soc. Am. 71, 679-688.

Zwicker, E. (1982) *Psychoakustik*, (Springer Verlag: Berlin).