PITCH MEASURING FROM SPECTRA OF NOISY SPEECH: AMPLITUDE THRESHOLDING VERSUS IDENTIFYING OF HARMONICS

V.Pikturna and A.Rudžionis

Speech Research Laboratory
Kaunas Technical University

ABSTRACT - Various criteria for identifying harmonic peaks in the FFT spectra are investigated. The low order linear prediction spectrum is used as an amplitude threshold crossing the upper parts of pitch harmonics

INTRODUCTION

One of the most important stages in pitch extracting algorithms is selection of pitch harmonics. In spectra of clean speech signals non-harmonic components are of low energy and can be separated by it from harmonic ones. In spectra of noisy speech more subtle methods for selecting pitch harmonics are necessary.The harmonics identifying methods can be divided into 3 groups: (1)identifying by amplitude of a spectral peak, (2)by its location and (3)by its shape.In the methods of the first group the maximum level of a spectral peak (Harris and Weiss,1963) as well as level relatively to neighbouring spectral peaks (Seneff,1978) is considered. A peak is identified as harmonic at once if its amplitude exceeds 18 dB (Sreenivas and Rao,1979). The additional requirements can be fixed concerning the amplitudes of peaks at different levels.Psychofysiological masking effect can also be used (Duifhnis et al,1982). The methods of the second group limit the minimal distance between harmonics. A peak is rejected if it stands too close a neighbouring peak of higher level. The accounting of the shape is based on the fact that it must correspond to the frequency response of the windowing function. The shape can be described by the width and the half-width of a peak (Sreenivas and Rao,1979).Similarity of a peak to the frequency response of a windowing function can be accounted with the help of the parabola approximating the upper 3 points of a spectral peak (Duifhnis et al,1982). The peak is harmonic when the approximation error does not exceed some value.

IDENTIFYING OF HARMONICS

The shape of a spectral peak

The visual inspection of FFT spectra shows that spectral peaks are of different sharpness, amplitude, symmetry.For describing sharpness and symmetry, we use parabola, approximating 5 upper points of a spectral peak (we suppose the 3 points approximation to describe a peak not enough accurately):

$$y_i=ai^2+bi+c, \qquad i=-2,-1,0,1,2 \qquad (1)$$

The coefficient a characterizes sharpness of a peak, symmetry being characterized by the approximation error:

$$\varepsilon=(1/5)\left[\sum_{-2}^{2}\left|ai^{2}+bi+c-sp(i)\right|\right], \qquad (2)$$

$sp(i)$ being the samples of the FFT spectrum in the zone of a local maximum.
The location of the maximum $sp(0)$ is determined more precisely taking the location of the maximum of the parabola into account:

$$i_{o}=-b/2a \; . \qquad (3)$$

Let δ be the distance between the maximum of the parabola and that of a harmonic.The success of approximation depends on the distance Δf between spectral samples and on the inter lay-out of a spectral peak and a harmonic. To learn the influence of these factors, the following experiments were carried out:

1) the frequency response $\mathcal{A}(e^{j\omega})$ of the Hamming window was calculated for every 1 Hz:

$$\mathcal{A}(e^{j\omega})=\int_{-\pi/\omega}^{\pi/\omega}(0.54+0.46\cos\Omega t)=$$

$$=0.54\,\frac{2\pi}{\Omega}\,\frac{\sin(\pi\frac{\omega}{\Omega})}{\pi\,\frac{\omega}{\Omega}}+0.46\,\frac{\pi}{\Omega}\left[\frac{\sin(\pi\frac{\Omega+\omega}{\Omega})}{\pi\,\frac{\Omega+\omega}{\Omega}}+\frac{\sin(\pi\frac{\Omega-\omega}{\Omega})}{\pi\,\frac{\Omega-\omega}{\Omega}}\right] \qquad (4)$$

$\omega=2\pi f$ being the running frequency, $\Omega=2\pi/T_{\omega}$, T_{ω} being the length of the analysis window (T_{ω} was fixed to the typical value of 30 ms).
For some fixed values of Δf (in the range from 5 to 30 Hz) the values of the parameters a, ε, δ were determined for the case when 5 spectral samples were symmetric in regard to the peak of the harmonic (4), and when shifted to 1,2,...,Δf Hz.

2) the same as 1) but introducing one more harmonic (4) in the neighbourhood of the examined one, at the distance of 80,100 and 200 Hz from it;

3) evaluating the parameters mentioned from the FFT spectra of synthesized vowels;

4) the same as 3), with natural speech.

In Fig.1 the distribution $\varepsilon(a)$ for natural speech is presented.

The shape of a spectral peak of noise

For producing noisy speech signal, the pseudorandom noise was generated with the values uniformly distributed in the range [0,1]. In the FFT spectra calculated from the short blocks of noise the energy dips are present which form spectral peaks.

To evaluate the shapes of peaks in the spectra of noise, the distribution $\varepsilon(a)$ was determined (Fig.2). Comparison of distributions shows both harmonic and noisy peaks to be equally symmetrical but harmonic peaks to be more sharp. The overlapping of the two distributions is too large for reliable identifying of harmonic peaks by their shape. A typical FFT spectrum of noise is presented in Fig.5.

The amplitude of a spectral peak

The amplitude H of a peak was calculated without using parabolic approximation (see e.g. Sreenivas,Rao,1978). The distribution of amplitudes was determined for the spectral peaks of natural speech signals (Fig.3) as well as for those of noise (Fig.4). The peaks in the FFT spectra of noise are evidently lower. The overlapping is however present.

AMPLITUDE THRESHOLDING

At high signal-to-noise-ratios (SNR), harmonic peaks rise high above the noise components, and a horizontal threshold can be efficient. It must be mentioned that the horizontal threshold cannot account for the shape of the vocal tract frequency response.Therefore either some harmonics are not evaluated (when the threshold is fixed high) or the non-harmonic components in the regions of formants are reached when the threshold is fixed low (Fig.6). The situation becomes essentially worse with the noisy signals when the spectral peaks of noise of large amplitude are present. In such situations a threshold is necessary crossing the most upper parts of pitch harmonics and going over the peaks of noise.

The linear prediction (LP) spectrum is known to follow the FFT spectrum when the both spectra are calculated from the same data block provided the parameters of the LP model are chosen correctly. After puting the LP spectrum on the FFT spectrum and shifting the LP spectrum properly downwards, it crosses pitch harmonics only, and usually all of them (see Fig.6). The main parameter responsible for the similarity of the shapes of the two spectra is the order of the LP model. Pitch frequency is determined in the narrow frequency range with one or two formants in it. The formant structure in such a range is represented sufficiently well by the 2nd-4th order LP model (Pikturna,1981). The 4th order model is however not suitable because it cannot be realized with the finite number of mathematical operations. Therefore the 3rd order LP model is used. The value of the shift depends on the SNR. The smaller SNR, the threshold must be fixed higher. We have experimentally determined the following values of shift: -5 dB, when SNR<10 dB, and -10 dB with cleaner signals.

When SNR is very bad, the speech sounds of low energy (mainly consonants) are fully destroyed:only spectral properties of noise are presented in their spectra.To build a pitch melody, the noise must be identified and ignored. For this purpose, we use the spectral flatness measure: the range of the LP spectrum. Our statistics shows the range of the 3rd order LP spectra of noise to be 0...11 dB while that of vocal sounds being 8...40 dB.

In Fig.7 and Fig.8 (male and female speaker correspondingly) the energies (*a*) and the pitch melodies of the German word "gepard" are shown: with clean (*b*) and with highly corrupted signal, without (*c*) and with (*d*) the spectral flatness measure.The peaks exceeding the LP threshold are in addition identified using parameters a, ε and H. After selecting harmonic peaks, pitch frequency is calculated as an average distance between them. Statistical one-step-ahead logic and the median smoother (Rabiner et al,1975) are used. Finally, the melody can be approximated by the 4th order polynomial. The so smoothed contour, without indicating unvoiced parts, is very suitable in teaching deaf persons to speak, in foreign languages studies, for pitch control in speech synthesizers etc.

CONCLUSIONS

Identifying pitch harmonics of noisy speech signals by their shape is not successful because the harmonics are similar to the peaks of the spectrum of noise. The essential improvement is reached by the special amplitude threshold i.e. the 3rd order linear prediction spectrum. It enables: (1)to fix the tops of the pitch harmonics and to ignore the noisy components; (2)to characterize the flatness of the FFT spectrum and to recognize reliably the noisy intervals. The melody contours maintain at SNR down to 0 dB, both for male and female speakers.

The method is realized with 64 spectral points in the frequency range 0-1430 Hz, 30 ms analysis window and 8 bit signal representation.

REFERENCES

Duifhnis,H.,Willems,L.F.,Sluyter,R.J.(1982)*Measurement of pitch in speech:an implementation of Goldstein's theory of pitch perception*,J.Acoust.Society of America,71,1568-1580.

Harris,C.M.,Weiss,M.R.(1963)*Pitch extraction by computer processing of high-resolution Fourier analysis data*,J.Acoust.Society of America,35,339-343.

Pikturna,V.(1981)*Investigation of speech formant analysis methods for μP-realization*,Doctoral dissertation,Kaunas.

Rabiner,L.R.,Sambur,M.R.,Schmidt,C.E.(1975)*Applications of a nonlinear smmothing algorithm to speech processing*,IEEE Trans.Acoust.,Speech and Signal Processing,ASSP-23,552-557.

Seneff,F.(1978)*Real-time harmonic pitch detector*,IEEE Trans.Acoust.,Speech and Signal Processing,ASSP-26,358-365.

Sreenivas,T.V.,Rao,P.V.S.(1979)*Pitch extraction from corrupted harmonics of the power spectrum*,J.Acoust.Society of America,65(1),223-228.

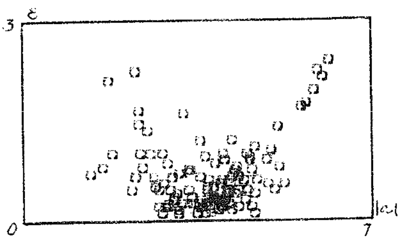


Fig.1.Symmetry/sharpness distribution (vowel /a:/)

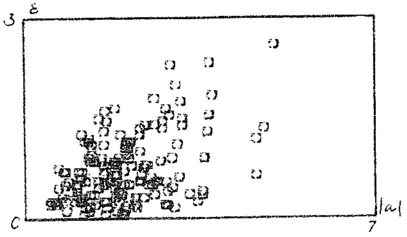


Fig.2.Symmetry/sharpness distribution (noise)

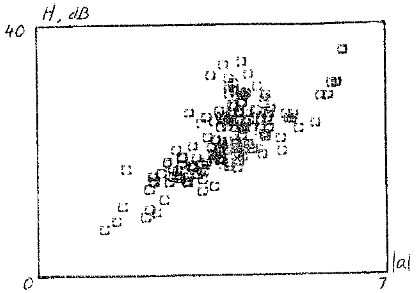


Fig.3.Amplitude/shrpness distribution (vowel /a:/)

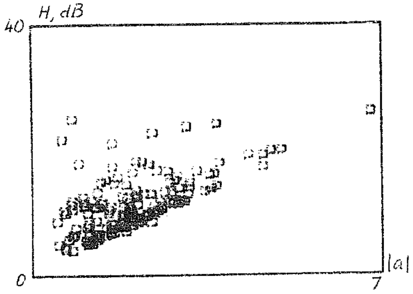


Fig.4.Amplitude/sharpness distribution (noise)

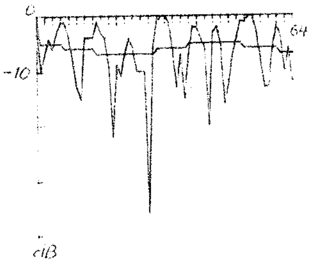
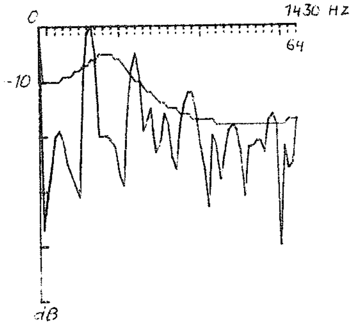


Fig.5.FFT and LP spectra (noise)



.Fig.6.FFT and LP spectra (female speaker)

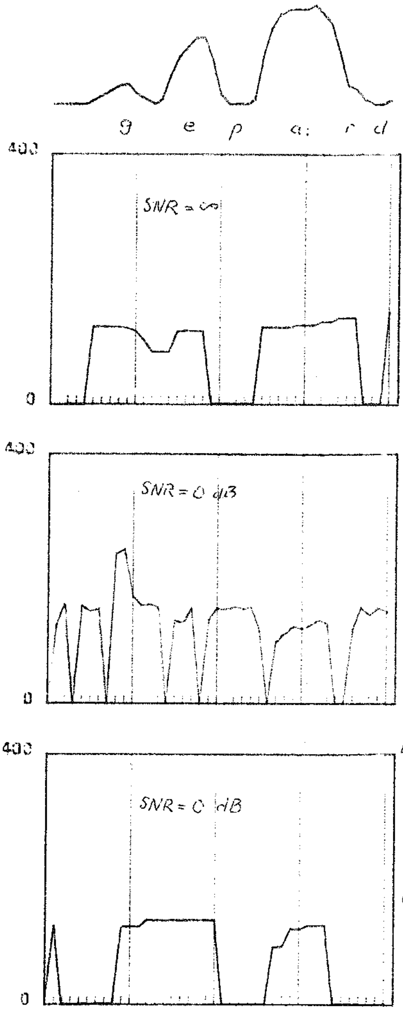


Fig.7.Energy and pitch contours (male speaker)

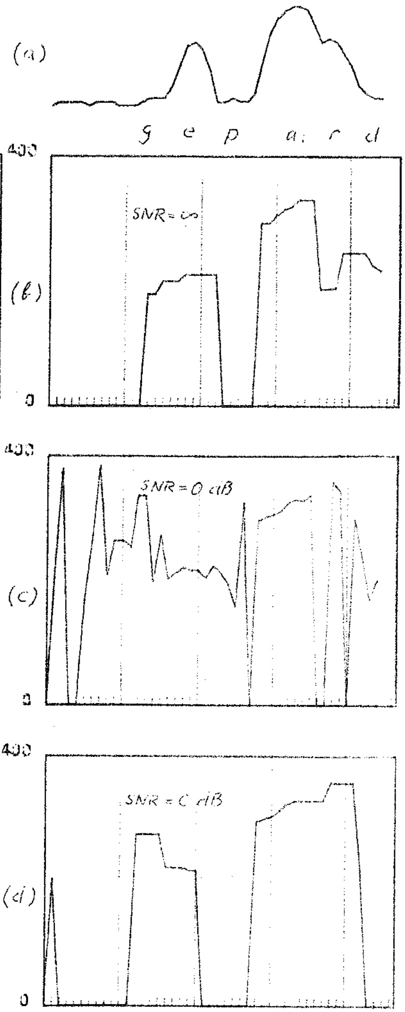


Fig.8.Energy and pitch contours (female speaker)