UNIVERSITY OF WEST BOHEMIA IN PILSEN,
DEPARTMENT OF CYBERNETICS

Some Experiments on Detection of Co-channel Speech

David Krivánka, Vlasta Radová

E-mail: dkrivank@students.zcu.cz, radova@kky.zcu.cz

1. INTRODUCTION

Co-channel speech occurs in situations when one speaker's speech is corrupted by another speaker's speech. Such situations occur for example in TV discussions, telephone calls, etc. Since the parts of speech signal containing multiple voices can cause problems in automatic speech or speaker recognition systems, it is reasonable to avoid the processing of such parts. In this paper we present some experiments with the co-channel speech detection based on so-called pitch prediction feature (PPF) [1].

2. COMPUTATION OF THE PITCH PREDICTION FEATURE

The principle of the PPF computation is based on a linear prediction (LP) model

$$s(n) = \sum_{k=1}^{Q} a_k s(n-k) + r(n) \,, \tag{1}$$

where $s(n)$ is the speech signal, $r(n)$ is the LP residual signal, a_k are the LP coefficients applied to the previous speech samples in estimating the current sample, and Q is the LP order. The residual signal $r(n)$ can be obtained from the speech signal $s(n)$ by applying a non-recursive filter $A(z)$

$$A(z) = 1 - F(z) = 1 - \sum_{k=1}^{Q} a_k z^{-k} \tag{2}$$

to the speech. The filter $A(z)$ is known as the prediction error filter and removes the near-sample correlations from the speech signal $s(n)$. The residual signal $r(n)$ contains thus mainly the distant-sample correlations caused by the pitch.

Let us now express the prediction error of the residual signal as

$$e(n) = r(n) - \beta_1 r(n-M+1) - \beta_2 r(n-M) - \beta_3 r(n-M-1) \,, \tag{3}$$

where M represents an expected value of the pitch period. Assuming a given value of M, the coefficients β_1, β_2, and β_3 are chosen in order to minimise the prediction error

$$E_{res}(M) = \sum_{n=0}^{N-1} e^2(n) \,, \tag{4}$$

where N is the number of samples in one frame. Using an autocorrelation method and denoting

$$\varphi(i,j) = \sum_{k=0}^{N-1} r(k-i)\, r(k-j) \tag{5}$$

the minimised $E^*_{res}(M)$ can be expressed as [1], [2]

$$E^*_{res}(M) = \varphi(0,0) - \frac{\varphi^2(0,M-1)}{\varphi(M-1,M-1)} - \frac{\varphi^2(0,M)}{\varphi(M,M)} - \frac{\varphi^2(0,M+1)}{\varphi(M+1,M+1)}. \tag{6}$$

Since the first term in (6) is independent of M, the optimal value of M can be found as that which maximises a function $u(M)$,

$$u(M) = \sum_{m=M-1}^{M+1} \frac{\varphi^2(0,m)}{\varphi(m,m)}. \tag{7}$$

The PPF is then defined as the standard deviation of the differences between the local peaks of $u(M)$, where $u(M)$ has been computed for $M \in \left\langle F_s/400, F_s/50 \right\rangle$, F_s is the sampling frequency. The local peaks are all peaks that are above the threshold given as 50% of the global maximum of $u(M)$ in our experiments.

3. EXPERIMENTAL RESULTS

In Figure 3.1, the signal $u(M)$ for a vowel pronounced by one speaker is depicted. The horizontal dotted line represents the threshold determining the local peaks. The PPF for this signal is 0.577. Figure 3.2 shows the signal $u(M)$ for a speech segment where two speakers are speaking together, both are pronouncing a vowel. In this case the PPF is about 32.067. It means speech segments containing speech of only one speaker has a low value of the PPF, whereas speech segments containing speech of more than one speaker are characterised with a high value of the PPF.

A problem, however, can occur with segments containing phonemes produced not only using the glottal pulses (like vowels) but also using a noise, e.g. like voiced fricatives. The signal $u(M)$ for the voiced fricative "Z"[1] pronounced by the same speaker as in Figure 3.1 is depicted in Figure 3.3 with a dashed line. The signal has higher peaks for those M that are near to the multiples of the expected pitch period, however, due to the noise component of the fricative there are not only one but several peaks overcrossing the threshold around the expected pitch period. The PPF for this signal is 47.869 that would cause the classification of the segment wrongly as containing the co-channel speech. For that reason we tried to use several types of filters on the residual signal $r(n)$. The best results were achieved using the Bessel low pass filter of the 6th order with the cut-off frequency 1 kHz. The resulting $u(M)$ signal

[1] The phoneme is marked using the SAMPA for Czech [3].

is depicted in Figure 3.3 with a solid line, the corresponding PPF is 0.707. It means the filter helps to classify the segment correctly as containing speech of only one speaker.

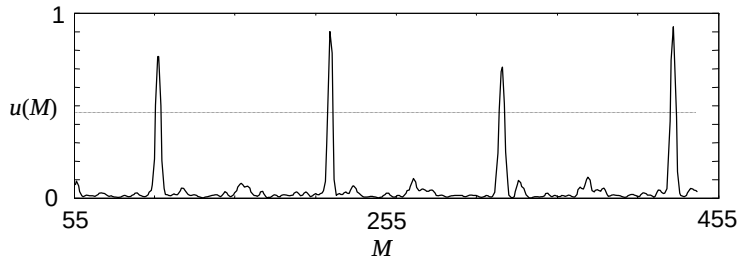

Fig. 3.1: Signal $u(M)$ of a single speaker

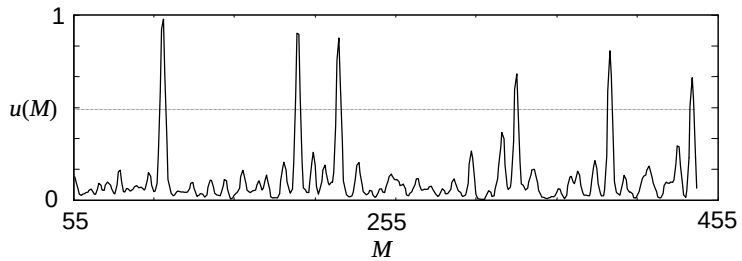

Fig. 3.2: Signal $u(M)$ of two speakers

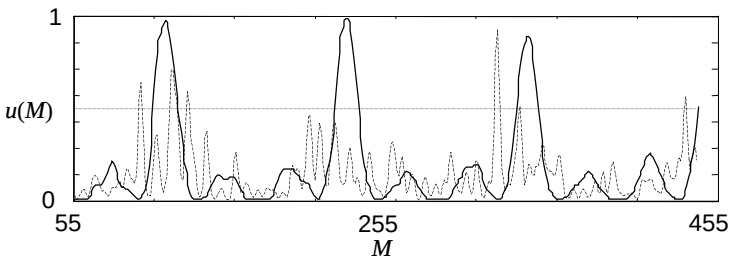

Fig. 3.3: Influence of the Bessel low-pass filter to the signal $u(M)$

4. CONCLUSION

A procedure allowing detection of those segments in a speech signal that contain the co-channel speech has been described in this paper. The procedure uses the so-called pitch period feature derived from a linear prediction model of speech. Results of experiments performed both for segments containing the co-channel speech and for segments not containing the co-channel speech has been presented.

ACKNOWLEDGEMENT

The work was supported by the Ministry of Education of the Czech Republic, projects no. F2286/03/G1 and MSM 235200004.

REFERENCES

[1] M. A. Lewis, R. P. Ramachandran, "Cochannel speaker count labelling based on the use of cepstral and pitch prediction derived features". *Pattern Recognition*, 34 (2001), pp. 499–507.

[2] J. Psutka, "The use of the LPC residual error autocorrelation to pitch period extraction". In: Proc. EUROSPEECH'89, Paris 1989.

[3] http://www.phon.ucl.ac.uk/home/sampa/