IDENTIFICATION OF A SPEAKER POSITION

A.E. Barabanov, A.Grusha
Saint Petersburg State University
Saint Petersburg, Russia
e-mail: Andrey.Barabanov@gmail.com

Abstract

A frequency and phase statistics are collected from the speech signal for estimation of the speaker position. The signal is recorded in four microphones. Frequencies are estimated by the multiband excitation unbiased criterion and the bell technique.

1 Introduction

A speech signal contains many important parameters of the sound wave that describe a speaker. One microphone receives a signal that say nothing about position of a speaker. But a set of microphones attached to a rigid body with a known geometrical configuration is sufficient for estimation of the 3 dimensional speaker position.

A necessity of a speaker position identification arises in the problem of a multivoiced speech separation. For instance, when a discussion of a number people with a music background is recorded a received signal is difficult to be processed by standard speech contraction algorithms from mobile telephone systems. If speaker positions do not move quickly they can be identified and the signal can be decomposed into a sum of voices.

2 The voiced signal parameters

A voiced signal of a speaker has many specific features in small and in large. Locally, during a couple of milliseconds the signal can be approximated by a periodic function in continuous time. This approximation changes slowly together with its intonation, allophone (like a letter) and volume.

The standard local model of a voiced speech signal is a sum of harmonics with multiple frequencies

$$s(t) = \sum_{k=0}^{N} A_k \cos\left(2\pi kFt + \phi_k\right)$$

where t is the continuous time in seconds, F is the Fundamental frequency in Herz, A_k and ϕ_k are the amplitude and the phase of the k-th harmonic.

After samplification with the sample rate F_s the sequence of the speech signal samples is described by

$$s_n = \sum_{k=0}^{N} A_k \cos\left(2\pi \frac{F}{F_s} kn + \phi_k\right).$$

This sequence is periodic if the Fundamental frequency F is an integer divisor of the sample rate F_s. Otherwise this model is only close to periodic. The Pitch period is defined as the period of the continuous function but it is measured in samples, $P = F_s/F$.

The most important parameters for the speaker location are the phases (ϕ_k). If a signal is measured by microphones located at a fixed points then the defferences of the phases for each harmonic is a measure of the distance difference between a speaker and the two microphones. This distance difference can be used for estimation of the three dimensional position of the speaker with respect to the rigid body with attached microphones.

3 Parameter estimation

A speech signal can be divided into a big number of small frames for Fourier analysis. In each frame the harmonic model can be studied and all its parameters can be estimated. Therefore, the signal provides a big amount of statistics for estimation of three values only - the coordinates of the speaker.

The problem of estimation of the Fundamental frequency has been studied for fifty years. A fast algorithm for identification of the integer value of the Pitch period [1]. It produces the unbiased estimate in the presense of a white noise with arbitrary signal/noise ratio.

In [2] a new approach was presented for accurate estimation of the Fundamental frequency by a sampled signal on a short interval. The approach was based on a special smoothing technique with extraction of spectrum of each harmonic in the full signal spectrum. The approach was called the "bell" technique.

Accuracy of the harmonic phase estimate is highly sensitive to accuracy of the frequency estimate. But it was shown that the bell technique provides sufficient accuracy for estimation of phases to accuracy of the same order as the frequency accuracy [3].

4 Position estimation

It can be easily proved that at least 4 microphones are necessary for 3 dimensional estimation of the speaker position by a signal with only one harmonic. The phase differences reduce the number of measurements to three, and the distances differences are determined up to the integer multiple of the wave length.

The problem of the position estimation is ill conditioned. Therefore, all estimates must be supplemented by their covariance matrices. The covariance matrices are obtained by linearization of the solution formulas around the nominal estimates. Influence of the signal nonstationarity and of the random noises were also studied.

5 Experiments

The experimental testbed was constructed with 4 microphones and one Web camera, see Fig.1. All signals from the microphones are synchronized with an Analog Digital Converter. The resulting 4 speech signals were processed by spectral analyser that gives estimates of Fundamental frequency and phases for all frames. Then three dimensional position was obtained by solution of an over system of an overdetermined nonlinear system.

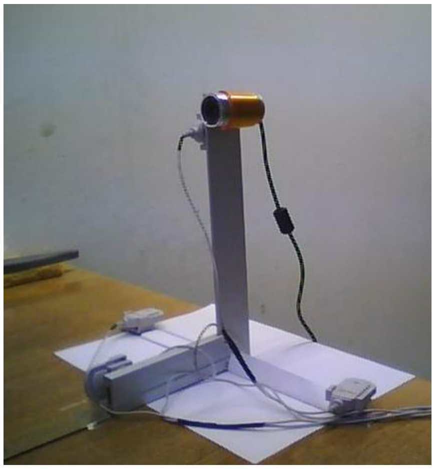

Fig. 1. The testbed.

We have studied the main harmonic of the speech signal. Each pare of microphones define a phase difference of this harmonic that determines a surfice in the space that must contain the signal source. this surface appeared to be a hyperboloid. An intersection of three surfaces gives an estimate of the speaker position, see Fig. 2.

Statistical analysis of the estimation accuracy was made.

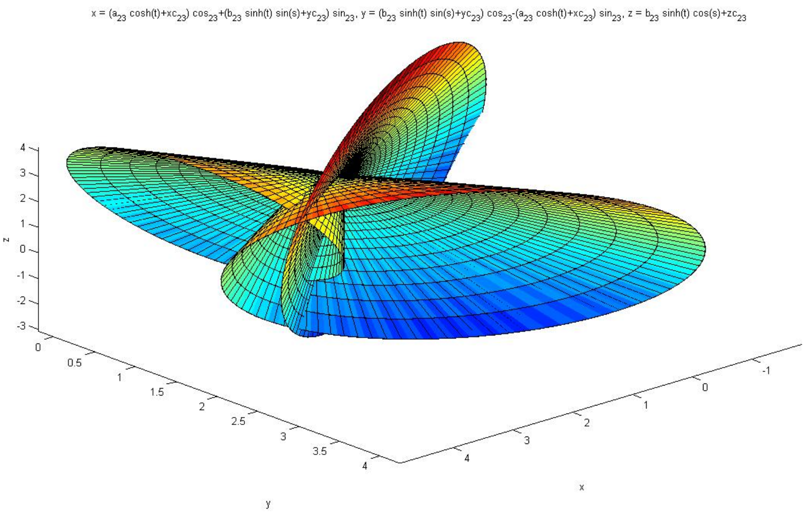

Fig. 2. Position estimation.

Conclusion

A speech signal was received by 4 microphones attache to a rigid body. It was modelled on short frames as a sum of pure harmonics with multiple frequencies. The model parameters were estimated by a combination of an unbiased criterion for integer Pitch detection and the smoothing bell technique.

The phase difference was used for the speaker position estimation. Statistical description of the final accuracy is provided.

References

[1] Daniel W. Griffin, Jae S. Lim. Multiband Excitation Vocoder. - IEEE Trans. on Acoustic, Speech and Signal Processing, v. 36, no. 8, August 1988, pp. 1223-1235.

[2] A.E.Barabanov, K.M.Putyakov, S.I.Salischev, V.I.Sitnikov. Echo compensation by equalizer with precise spectrum estimation. The 21st AES International conference "Architectural acoustics and sound reinforcement", St. Petersburg, June 1-3, 2002, pp. 357–362.

[3] A.E.Barabanov, D.V.Romaev. Adaptive filtering of tracking camera data and onboard sensors for a small helicopter autopilot. Proc. of the 3rd IEEE Multi-conference on Systems and Control. 2009, Saint Petersburg. July 8-10.