Citation:

ILLÉNYI, A., WERSÉNYI, GY.,
*Averaged speech signal samples generated by speech chorus method.*

Proceedings of the International Békésy Centenary Conference on Hearing and related Sciences, 1999, Budapest, pp. **115-120**.

# AVERAGED SPEECH SIGNAL SAMPLES GENERATED BY SPEECH-CHORUS METHOD

Wersényi, Gy. Ph.D. student, wersenyi@sparc.core.hu
Illényi, A. Ph.D. consultant, illenyi@sparc.core.hu
Georg Békésy Acoustic Res. Lab., Dep. of Telecommunication and Telematics, TU of Budapest

## Introduction

Békésy observed that different kinds of signals described in the time space cause the same perception. E.g. if we are listening to a vowel from a different distance and/or a different angle of incidence we are able to recognize it though these signals differ from each other a lot (on the oscilloscope). This problem is well known in the area of the acoustical information transmission.

We also know that the human body (outer ears, head and torso) influences the signals. The Head-Related Transfer Functions (HRTF) are complex transfer functions which are mainly describing this effect in the frequency space. The HRTF measurement system we installed and the influence of the everyday life environment is handled in [2].

During this investigation a huge speech database (BABEL) was recorded for other applications. We are going to use this for creating input signals for our measurement system in the future. In this paper we will check and scrutinise the spectral properties of these samples and examine the speech-chorus signals (which were created to include all the spectral properties of human speech in one short sample file).

## HRTF measurement system using speech input signals from the database

This database was originally created for language independent speech recognition systems (for telecommunication systems). After the segmentation it is used to teach the neural networks. The recordings were made in the anechoic chamber using one channel, 20 kHz sample frequency and 16 bit. The average SNR is 54 dB [1]. The speakers are both male and female, younger and older persons, children, and children with hearing aids as well. They speak different but specially created Hungarian texts. English and German samples were also used for examination.

As presented in paper [2], we installed a DSP based HRTF measurement system in the anechoic chamber. We measured the head and torso simulator's

complex HRTF's with an average SNR of 86 dB using pseudo-random white noise. We showed [that] everyday life environment (clothing, hair etc.) do influence the HRTF's and we presented some informative results of the most significant changes.

At the moment we are at an intermediate station. Our goal is to create special speech input signals for this measurement system instead of the noise. We try to make an „averaged" signal of the human speech using the database mentioned above. We took different samples and mixed them up to each other. In the end there are 20 sec. long speech-chorus sample files from various groups of subjects. We investigated these files spectrally using 16384 points FFT and Blackman-Harris windowing.

The different FFT window types will give different frequency graphs. The Triangular window gives a more precise frequency estimate, but is also the noisiest, which means other frequencies will be shown as present, even though they may be much lower in volume. At the other extreme, the Blackmann-Harris window has a more broad frequency band which is not as precise, but the sidelobes are very low, making it easier to pick out the major frequency components [5].

Under examination we also found that the signals should exceed 5-6 sec of time to be good enough for a „speech" signal (it is not noisy spectrally anymore).

A control measurement of the spectrum of the „silence" between the sentences in a random sample file showed the average signal-to-noise ratio in the frequency caused by the recording and the signal processing (FFT). It is quite constant and has a value of -100 -110 dB.

## The speech-chorus signals

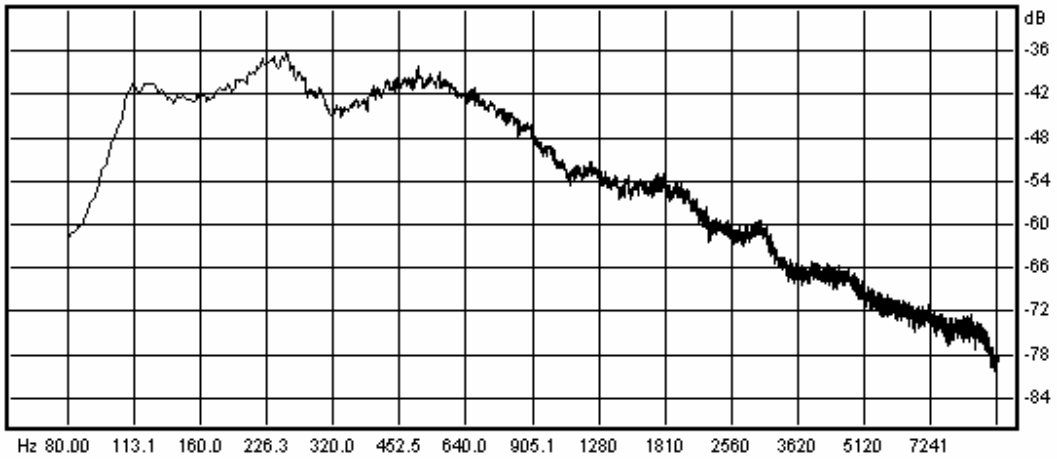


Fig.1. Hungarian sample file, 15 male and 15 female subjects ensemble (20-68 ages).

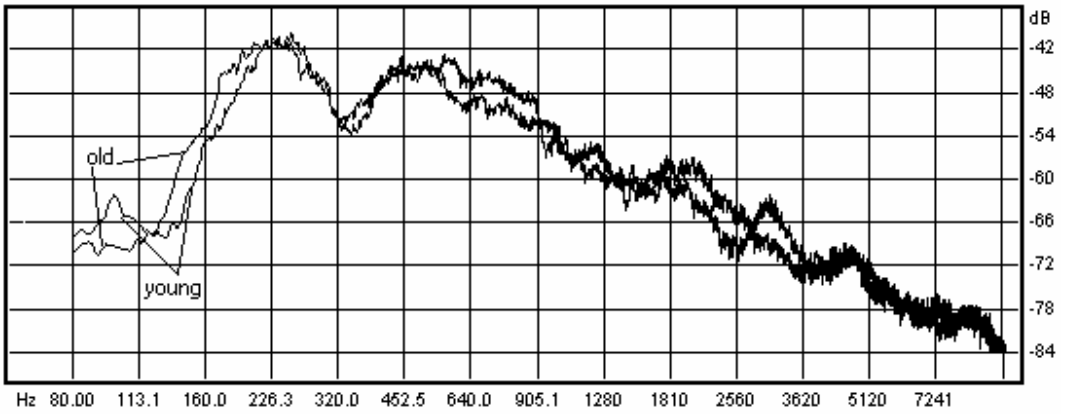


Fig.2. Hungarian sample files, 8 younger (20-32-year-old) and 7 older (44-67-year-old) females. It is to be seen that no significant differences appear. We are not able to decide the age of the speaker from a recording.

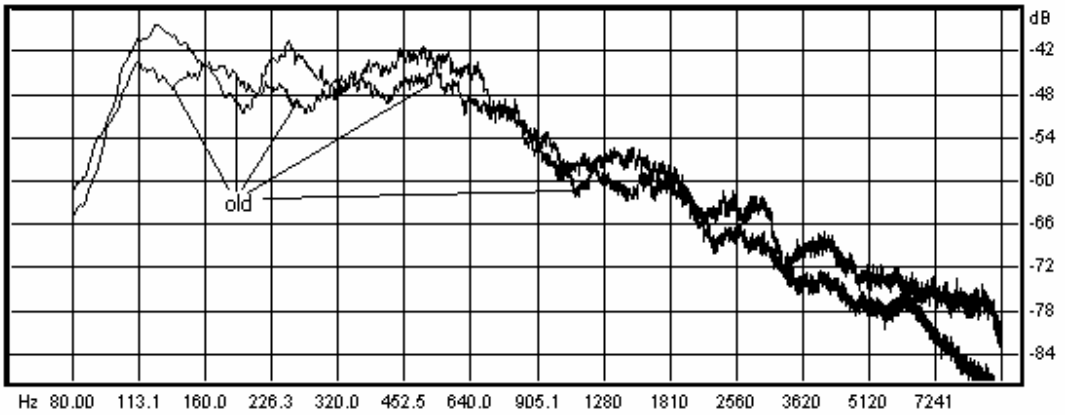


Fig.3. Hungarian sample files, 8 younger (21-37-year-old) and 7 older (44-68-year-old) males. There is no significant difference, but the curve of the older speakers is more linear.

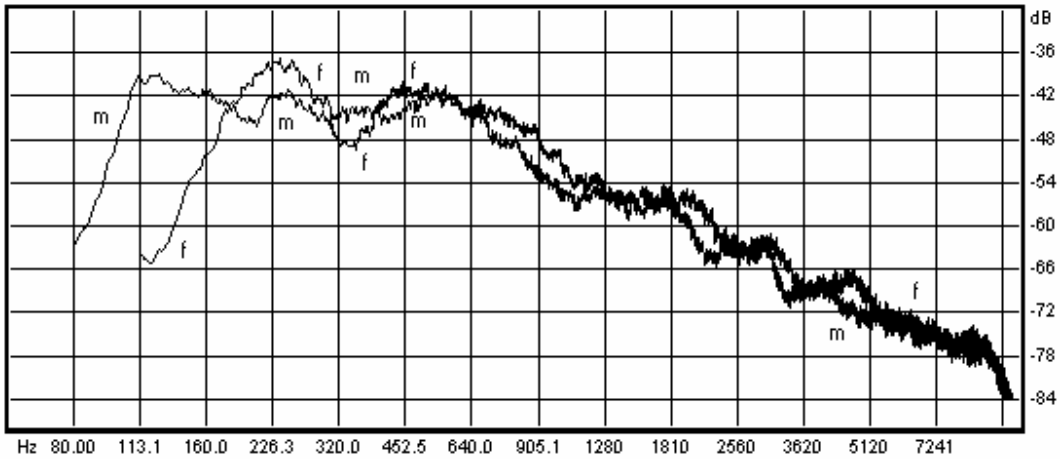


Fig.4. Hungarian sample files, 15 males and 15 females for comparison. This figure is practically the same as the one measured by Tarnóczy [3]. The typical waviness of the female curve between 160-540 Hz is based on the nearness of the pitch frequency and the $F_1$ formant frequency. In male subjects this distance is bigger so the curve is more linear. We can select a speech from a recording by many speakers based only on the pitch frequency.

On closer examination we found that there is no significant difference between the spectrum of male and female speakers in 9-year-old children. The difference between the genders comes up after the change of the voice in male subjects.

The same is observed in 11-year-old children, and there is no significant change in the 9-year-old speakers though the older ones are speaking more clearly. The younger children show a bigger dynamic range mainly in the lower frequency range (160-1000 Hz).

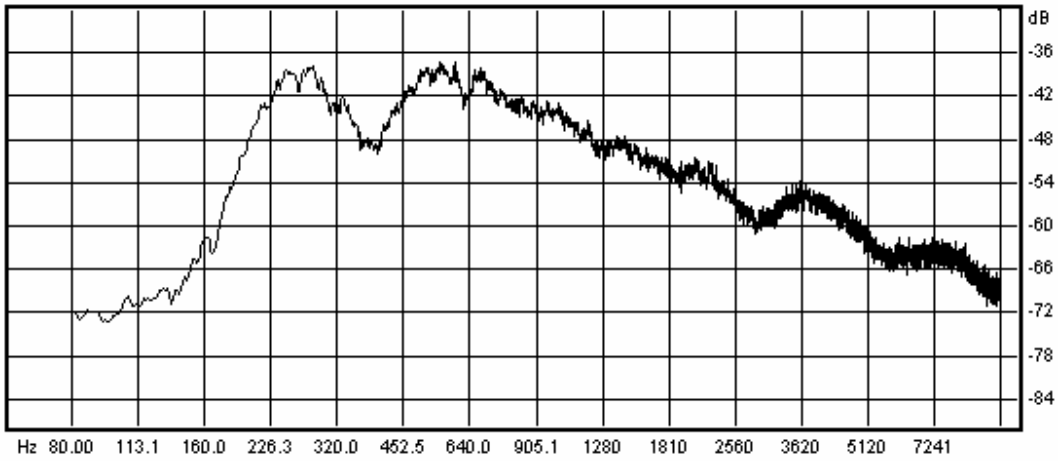


Fig.5. Spectrum of healthy Hungarian children (ensemble). 14 speakers between 5-11 years of age of both genders. The FFT was made from a 13 sec sample file.

As expected, adults have more components at lower frequencies (80-226 Hz) than children. The dump at 300 Hz by adults is about 6 dB, the same by children is 12 dB and it shifts up to 400 Hz (fig.1. and fig.5.).

The growth with age is noticeable spectrally as well. In male speakers the 300 Hz resonance disappears with the age (decrease about 10 dB) but on the other hand the dump at 400 Hz is rising. This causes a bigger „linearity" in older speakers. In female speakers there is not such a big change. The character of the curve stays the same. The resonance at 200 Hz does not vary at all. Adult female spectrum begins at ca. 110 Hz, that of the children at about 200 Hz. The 6 dB subsidence of the 300 Hz dump is observed.

Because of the fact that between the male and the female children no big differences appeared we can say that male subjects are going through stronger voice mutation spectrally (change of voice).

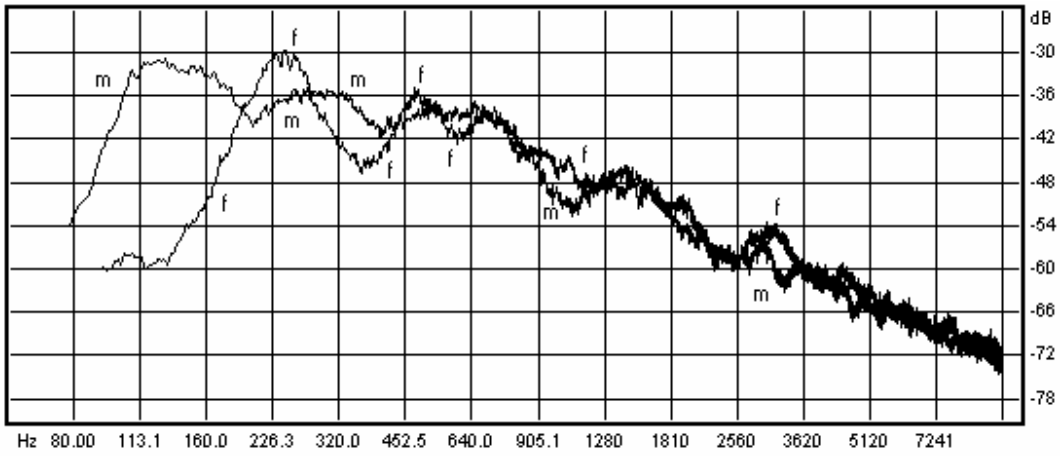


Fig.6. German sample files (15 males and 15 females). Fig.7. is the ensemble. Compare with fig.4.

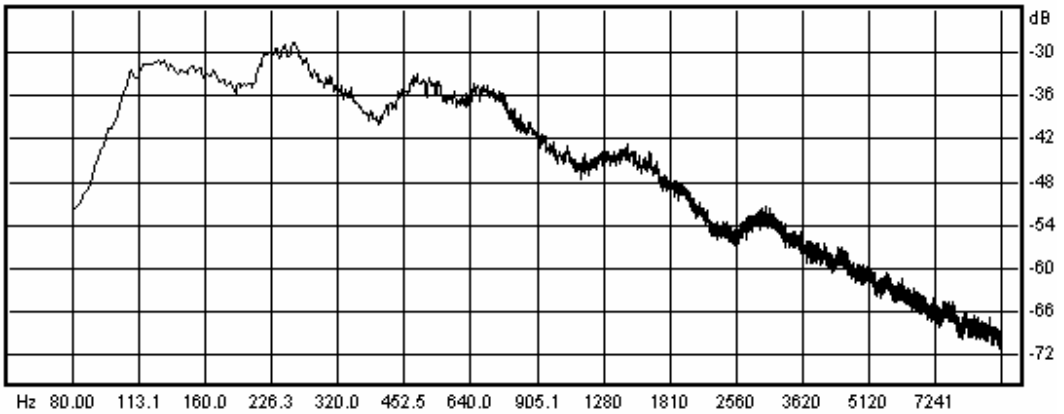


Fig.7. German sample file, 15 male and 15 female ensemble. Compare with fig.1. and fig. 8.

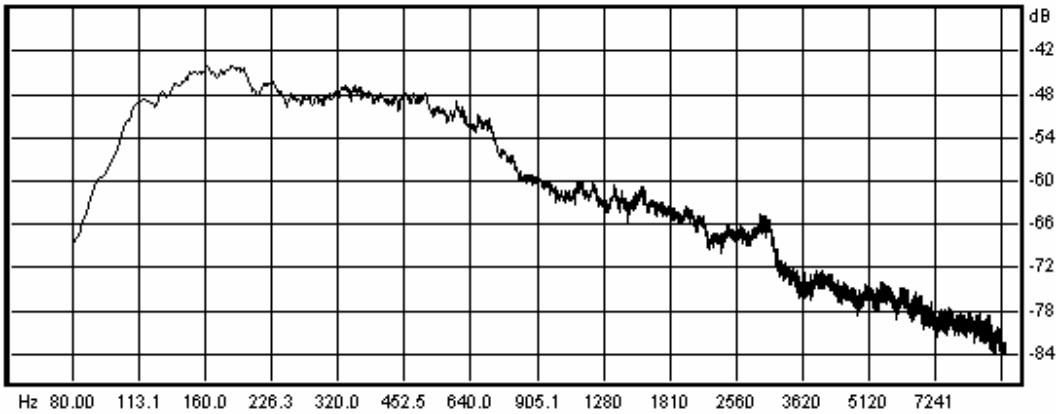


Fig.8. English sample file for comparison (6 male and 1 female ensemble).

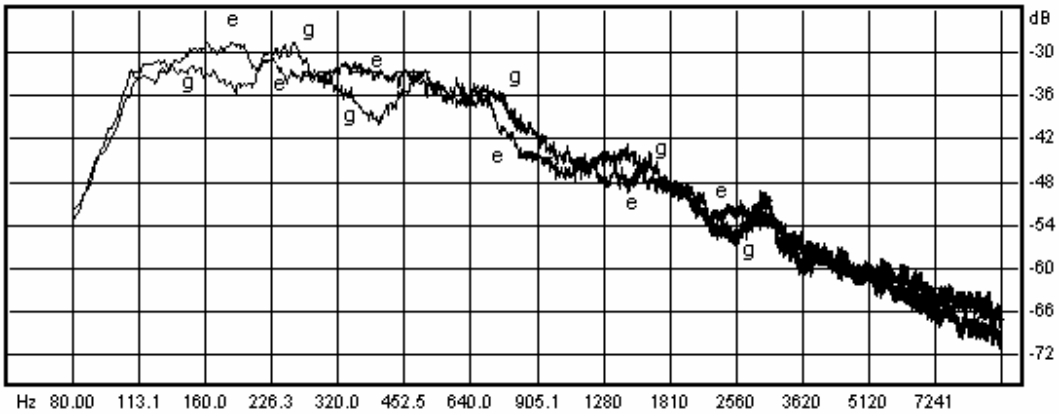

Fig.9. English and German ensembles in one figure for comparison (fig.7. and fig.8.). Because of the difference between the recordings levels the reference is the maximum of the spectras. The English spectra is more linear and has attenuation between 150-220 Hz and 320-450 Hz, and dumping at 300 Hz. Similar effect is observed by the English-Hungarian comparison. By German-Hungarian comparison is only a little dumping effect at 200 Hz and between 320-640 Hz in the German spectra.

## Conclusion

Our main field of interests is the acoustical information transmission and „decoding" from the acoustical signals. Speech-chorus signals were rarely used so far with a few results.

There are now various length of sample files from both gender in different age-groups and from different languages using the BABEL database. We are going to use this as an „averaged speech signal" for the HRTF measurement system. To make sure of the correctness, these signals were spectrally investigated and compared with each other and with former results.

Maybe they can be useful for people working with telecommunications and telematics having all the figures all together.

### References

[1] Vicsi, K., Vig, A.*"Az elsõ magyar nyelvû beszédadatbázis"*, Beszédkutatés '98-Beszéd, spontán beszéd, beszédkommunikáció 163-177 o., Budapest, 1998.
[2] Illényi, A., Wersényi, Gy.*"Discrepant in binaural tests and senses of soundfield parameters"*, G. Békésy Anniversary Conference, 1999, Budapest.
[3] Tarnóczy féle ábra??
[4] Vicsi, K., Csatári, F., Bakcsi Zs. *„Distance score evaluation of the visualised speech spectra at audio-visual articulation training"*, EUROSPEECH'99 Conference Proceedings, Budapest, 1999. (Forthcoming paper)
[5] Help file of Cool Edit '96 (Syntrillium Software).