# DATABASE DESIGN METHOD FOR SPEECH SYNTHESIZERS APPLIED TO THE REHABILITATION OF DISABLED PEOPLE

K. Lukaszewicz* and M. Karjalainen**

*Institute of the Biocybernetics and Biomedical Engineering PAS. ul. Ks. Trojdena 4 Warsaw Poland
email: konrad.lukaszewicz@ibib.waw.pl

**Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Otakaari 5A Espoo Finland
email: matti.karjalainen@tkk.fi

**The speech synthesizer can be used as an electronic rehabilitation device for the speech disabled people. To obtain the highly desirable individual voice, an appropriate speech sample database has to be collected. This paper presents a simple method of determining the voice similarity by analyzing a set of very short sounds. It can be used to shorten the time needed for database collection by selecting a ready-made database, most similar to the patient's voice. The voice similarity was defined as the sum of single elementary similarities of short sound pairs. To find the similarity between voices, a large number of pitch-length sounds were extracted from natural voice signals from different realizations of open vowels 'a' and 'o'. This method is oriented to the use of the microphonemic speech synthesis based on waveform concatenation. This simple and computationally low-cost speech synthesis method can be applied in small portable devices and used for the rehabilitation of speech disabled people.**

## 1. Introduction

Speech is the most natural way for communication. Disabled people who cannot speak can still provide conversation using text-to-speech synthesis (TTS) [1][2][3]. The speech synthesizers available today are of so high quality that the artificial speech can be used for oral communication. The new methods of speech synthesis give possibility to obtain synthetic speech similar to a particular subject's voice. It is important when synthetic speech is used for helping the disabled people who loose their natural voice. Usually each application has its own needs [4], and the expected quality necessary for a particular application can vary. Nowadays, the largest group of people, who are using synthetic speech daily, is the blind community. Usually they use speech for the communication with computer-user interface. Then the quality of speech is not so important as its flexibility and quick response. For people who have lost their voice, lack of verbal communication becomes one of the most important problems. The speech synthesizer can in this case be used as speech prosthesis. Synthetic speech used by speech disabled people has to fit individual voice characteristics. It is not enough to have a female voice for women and a male voice for men but also the speech should have personal features of the particular user. This requires the collection of a specific database for each individual person. Currently the time and effort that is needed to collect a full set of required data for speech synthesis is too long to be applied for each patient separately. On the other hand the speech unit has to be of portable size to allow people use it outside buildings, this pointing to speech synthesis methods of low computational load.

The microphonemic method of speech synthesis [5][6][7] belongs to the group of concatenative speech synthesizers. The synthetic speech signal is generated by concatenation of short (typically pitch period size) waveform elements. Modification of the length and size of each sound element and smoothly connecting them can generate high quality speech signals where the individual speech features are maintained. The database used for the microphonemic speech synthesizer consists of a large number of sound elements, which represent all the diphonemes in the particular language. The length of each element is typically a pitch period.

One of the main problems related to designing a microphonemic speech synthesizer is the collection of the necessary sound set. To limit the number of data elements and time needed to collect the whole set, the number of different realizations of each phoneme has to be defined. The higher the number covered by the database, the better the quality of synthetic speech that can be obtained, but this will increase the size of the database and the time needed to collect it. One possible way to limit the time required for collecting data for a particular user is to apply a ready-made part of a database with addition of some elements coming up with the particular voice features. The addition of short personally oriented elements to the main database takes a shorter time than the collection of a full set of sound elements for each patient.

To select the most suitable ready-made parts of a database for a particular voice, a similarity measure has to be defined. The similarity used here has been determined by analyzing the auditory spectrum of short elementary sounds. In our analysis the stability of articulation of the Polish vowels ‘a’ and 'o' was assessed. A sound element of single pitch period length has been taken from the middle part of each realization of the vowel. Totally 346 short sound elements of vowel 'a' and 215 of vowel 'o' were collected for each voice. Four different voices were tested.

The algorithm of sound similarity presented here can be used to determine the number of substantially different realizations of each vowel for the natural voice. While a few databases were collected and are ready for the synthesis of different voices, a new particular voice can be obtained by modifying an existing database. This algorithm can be used for choosing the most suitable sample for the modification database. Modification of the database which includes vowels most similar to a particular voice could give better quality synthetic speech.

## 2. Auditory listening tests

The elementary (pitch length) sounds were collected from 47 different words of a speaker. The Polish vowel 'a' was tested. The extracting point was placed at the middle point of the vowel realization (Figure 1). The beginning and the end of the selected signal were placed at zero crossing points (Figure 2) to obtain a smooth connection in the concatenation. The collected elements were standardized in length and amplitude by multiplying with a trapezoidal window for increasing the pitch or by adding a short silence for decreasing the pitch. So all played sounds had the same length and amplitude.

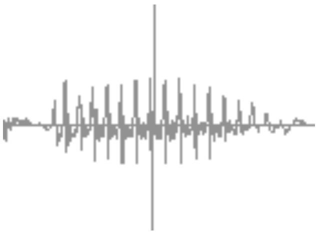

Figure 1. *Extraction position.*

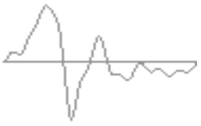

Figure 2. *Extracted sound.*

By concatenating each stored element 40 times, 47 different sounds were generated so that the duration of each sound was about 0.3 sec. To test the similarity, pairs of two sounds were played and compared. The first sound was always the same while the second was changed.

In the auditory test for each listener, 47 pairs were played. The listener could hear each pair as many times as he needed in order to make the decision whether the played sounds were similar or not, i.e., a binary decision. 22 tests were performed by 7 different subjects. The first pair was made of two identical sounds, thus in each test the result for comparison of the first pair was always positive.

Some sounds were determined by the listeners as being more similar to the source sound (sound number 1) than the others. If we look at the histogram of similarity (Figure 3) we can see that the sounds 6, 11, 20, 33 and 45 are the most similar ones. On the vertical axis the number of positive decisions was placed, while the horizontal axis describes the sound number that was compared with the sound number 1.

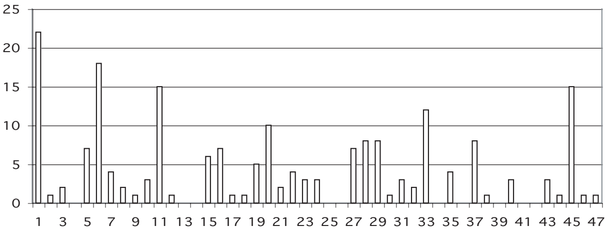

Figure 3. *Auditory test results as a histogram.*

## 3. Comparison by auditory spectrum

Even when the elementary sounds were extracted from utterances of one speaker and very similar context (the sounds were extracted from the same tri-phone CVC), the listeners noticed significant differences. The perceptual differences can be computationally estimated by using a proper auditory model. We have applied auditory spectrum that warps the Fourier spectrum to the Bark scale and the magnitude scale to loudness density (also called specific loudness) [7][8].

Considering the auditory spectra of similar and different pairs we can see large differences in the spectral patterns. Figures 4 and 5 present the auditory spectra of two sound pairs. Figure 4 shows the auditory spectra of two similar sounds and Figure 5 presents auditory spectra of two different sounds.

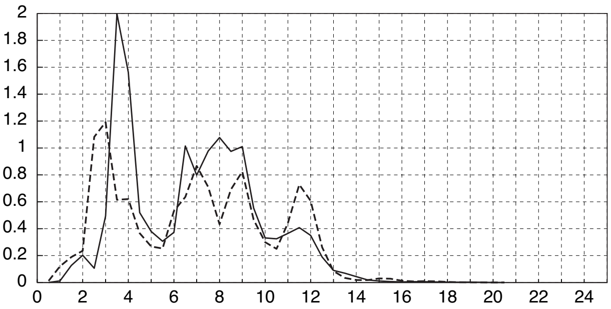

Figure 4. *Auditory spectra of two similar sounds: x-axis: pitch (Bark), y-axis: loudness density (sone).*

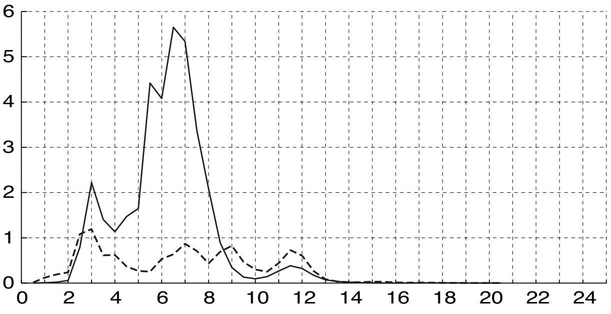

Figure 5. *Auditory spectra of two different sounds: x-axis: pitch (Bark), y-axis: loudness density (sone).*

Selected by the listeners, similar sounds had a similar layout of the auditory spectrum. The number of energy bands (see definition below) was the same and the position and the energy in each band was quite similar.

## 4. Similarity determination algorithm

For finding the similarity of two elementary sounds, their auditory spectra were considered. For comparing the pattern of energy in the spectrum a set of 17 parameters was defined: number of the energy bands, the total loudness in each band, the position of each energy band and the center of [9] gravity for each band. To define the beginning and the end of an energy band, the positions of local minima were searched for. Related to auditory resolution, a minimum in auditory spectrum was searched for as any range of drop in energy that is wider than 2 Barks on the critical band scale [9]. The width of a minimum region was then set to 1 Bark around the minimum, see an example in Figure 6.

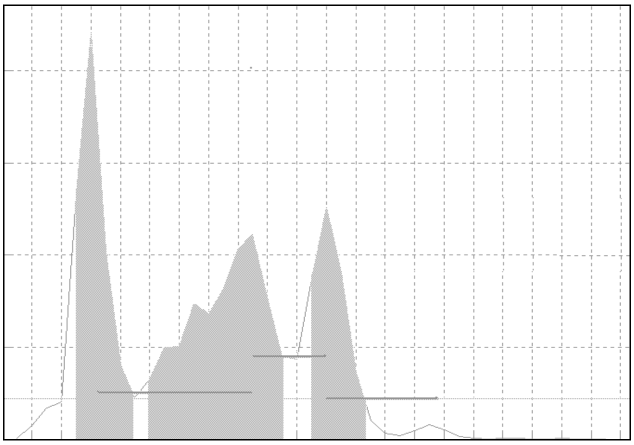

Figure 6. *An example of found bands and minima in an auditory spectrum (on Bark vs. sone scale).*

The position of the center of gravity for each band was defined as the first moment calculated for each band separately.

These parameters were used to compare the auditory spectra of the sound pairs. The sounds were similar if: the numbers of bands were the same, the positions of the bands were at the same region, the total loudness of each band was at the same level and the positions of the centers of gravity were at the same frequency regions. The band position regions, the loudness level and the centers of gravity frequency regions were defined based on the (described above) auditory tests results.

## 5. Voice similarity

The algorithm defined above was used to test the similarity between four different voices: TOM, KAS, HAN and KAM. From each voice, 346 microphonemes from vowel 'a' (from the middle part of the vowel) and 215 microphonemes from vowel 'o' (from the middle part of the vowel) were collected. The similarity of voice pairs was calculated as the sum of the single tests results. In each case, the similarity of a short sound pair was calculated. The result of a test was 1 when the sounds in the pair were similar and 0 if the sounds in the pair were different. In each pair of two short signals the first one was taken from the first voice and the second one from the second voice. So for each pair of voices the total number of tests was 346x346 = 119025 for vowel 'a' and 215x215 = 46225 for vowel 'o'. The numbers of positive result of the single tests for each pair of voices were placed at Tables 2 and 3.

Table 1. *Similarity of voice pairs for vowel 'a'.*

| **A** | TOM | KAS | HAN | KAM |
|---|---|---|---|---|
| TOM | 2994 | 0 | 73 | 1 |
| KAS | 0 | 790 | 9 | 167 |
| HAN | 73 | 9 | 1006 | 59 |
| KAM | 1 | 167 | 59 | 1028 |

Table 2. *Similarity of voice pairs for vowel 'o'.*

| **O** | TOM | KAS | HAN | KAM |
|---|---|---|---|---|
| TOM | 3463 | 1 | 312 | 8 |
| KAS | 1 | 689 | 6 | 342 |
| HAN | 312 | 6 | 1765 | 5 |
| KAM | 8 | 342 | 5 | 909 |

It can be noticed that a higher similarity result was obtained for pairs of the same voices. For example voice pairs TOM – TOM gave result 2994 for vowel 'a' and 3463 for vowel 'o'. For both vowels the most similar speaker pair of tested voices was KAM-KAS and the most different one TOM-KAS.

Considering each voice we can notice that analyzing the vowel 'a' and vowel 'o' we have obtained similar results. The most similar voice for voice TOM was HAN and the next similar voices were KAM and KAS. The same order of voice similarity was obtained when the vowel 'o' was considered.

## 6. Parameter histograms

Each one of the recorded voices comes from a different speaker. Considering some of these defined above parameters we can notice that the vowel realizations can be very different even for a single speaker.

In figures 7 and 8, histograms of the numbers of bands detected for four different voices are presented. That information can be used for constructing the database for the microphonemic voice synthesizer. For example for speaker HAN the vowel 'a' in most cases has three or two bands. Those realizations of vowel 'a' could be determined as representative ones for this particular voice.

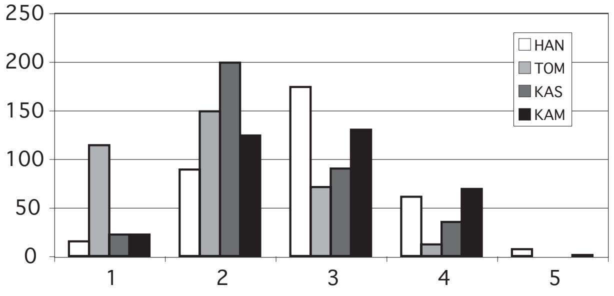

Figure 7. *Histogram of numbers of bands (horizontal axis) in vowel 'a' for four different voices (occurrence on the vertical axis).*

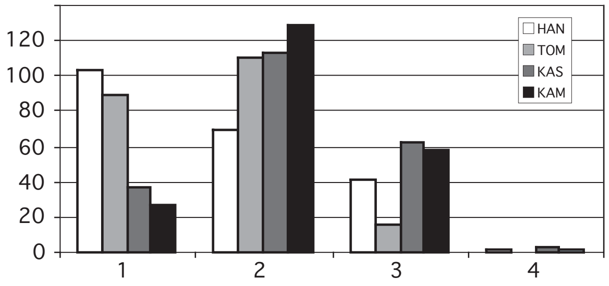

Figure 8. *Histogram of numbers of bands (horizontal axis) in vowel 'o' for four different voices (occurrence on the vertical axis).*

## 7. Discussion

Using this simple method of sound comparison we can notice how many different realizations of each vowel we can find in a particular voice. The feature histograms show the most common forms of each vowel for a particular voice. It can be noticed that for example in the voice HAN the most common number of energy bands for vowel 'a' was 3 while for voice KAS it was 2 (Figure 7). By analyzing the auditory spectrum we can determine how stable a voice is and how similar it is to another one. According to the tests it looks like for one speaker each of the tested vowels (e.g. 'a' and 'o') can be represented by two different realizations. In the tested voices it covered 80% of all realizations of vowels 'a' and 'o'.

The voice similarity algorithm could be used to complete a generic database, which can be used for synthesis of a few similar voices. The database for one voice could be modified by the exchange of vowels and some voiced consonants to obtain another individual sounding voice. In this case the most significant (for a particular speaker) realizations of each vowel have to be applied as the database modification elements.

## 8. Conclusion

The presented results show that it is possible to compare the similarity of two voices by analyzing auditory spectra of sets of short signals extracted from many different points of a single subject’s utterance. The presented method of analyzing voices requires low computational effort and is based on pitch period sized sound signals that are typical for the microphonemic method. This helps in efficient generation of databases for the microphonemic synthesis of individual sounding voice synthesis for speech disabled subjects.

## 9. References

[1] Karjalainen M., Laine U., Toivonen R. (1980). "Aids for the Handicapped Based on SYNTE 2 Speech Synthesizer". Proceedings of IEEE ICASSP 80, (3): 851-854.

[2] Wloskowicz D., Lukaszewicz K., Radecki K., "Implementation of Synthetic Speech in a Phone Communication System for Deaf-mute People". Polish J. Med. Phys. & Eng. 1999, Vol. 5, No. 1 (15), Warszawa 1999 p. 33-39.

[3] Wloskowicz D., Lukaszewicz K., Radecki K., "Phone communication by means of synthetic speech for deaf-mute people". ESEM 2001 International Journal of Health Care Engineering, Technology and Health Care Volume 9, Numbers 1,2,2001 Belfast, May 2001. Proceedings p. 58-60.

[4] Lukaszewicz K., "The Ultrasound Image of the Tongue Surface as Input for Man/Machine Interface". INTERACT '03 Human-Computer Interaction. Zurich, September 2003. Proceedings p. 825-828.

[5] Lehtinen L., Karjalainen M., "Individual Sounding Speech Synthesis by Rule Using the Microphonemic Method". Proceedings of Eurospeech 89 (2): 180-183

[6] Lukaszewicz K., Karjalainen M., "Microphonemic Mathod of Speech Synthesis". Proc. IEEE ICASSP 87, Dallas 1987 (3): 1426-1429.

[7] Lukaszewicz K., Karjalainen M., "Microphonemics – High Quality Speech Synthesis by Waveform Concatenation". ICPhS, Tallinn 1986 (6): 48-51.

[8] Karjalainen M., "A New Auditory Model For The Evaluation of Sound Quality of Audio Systems". Proceedings of IEEE ICASSP 85: 608-611.

[9] Zwicker E. and Fastl H., Psychoacoustics – Facts and Models. Springer Verlag, 1990.

[10] Christovich L. A., "Central Auditory Processing of Peripheral Vowel Spectra". J. Acoust. Soc. A. 77 (3), March 1985: 789-805.

[11] Hunt A., Black A. (1996). "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database". Proceedings of IEEE ICASSP 96: 373-376.

[12] Hon H., Acero A., Huang X., Liu J., Plumpe M., "Automatic Generation of Synthesis Units For Trainable Text-To-Speech Systems". Proceeding of ICASSP 98

[13] Charpentier F., Moulines E. "Pitch-Synchronous Waveform Prosessing Techniques for Text-to-Speech Synthesis Using Diphones". Proceedings of Eurospeech 89 (2): 13-19.