Simple Method of Determining the Voice Similarity and Stability by Analyzing a Set of Very Short Sounds

Konrad Lukaszewicz[1] and Matti Karjalainen[2]

[1]Institute of the Biocybernetics and Biomedical Engineering PAS.
ul. Ks. Trojdena 4, Warsaw, Poland
konrad.lukaszewicz@ibib.waw.pl

[2]Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5 A Espoo, FI-02015 TKK, Finland
matti.karjalainen@tkk.fi

Abstract. This paper presents a simple method of determining the voice similarity by analyzing a set of very short sounds. A large number of pitch-length sounds were extracted from natural voice signals from different realizations of open vowels 'a' and 'o'. The voice similarity was defined as the sum of single elementary similarities of short sound pairs. This method is oriented to the microphonemic speech synthesis based on waveform concatenation, and it could help to limit the time needed for database collection. This simple and low computational load speech synthesis method can be applied in small portable devices and used for the rehabilitation of speech disabled people.

1 Introduction

Many available speech synthesizers can produce very natural voice. The quality of voice cannot however be considered without relationship to application. Usually each application has its own needs, and the expected quality necessary for a particular application can vary [2,6]. 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, the verbal communication becomes one of the most important problems. The speech synthesizer can in this case be used as speech prosthesis. The synthetic speech used by speech disabled people has to fit individual voice characteristics. This demand needs 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.

The microphonemic method of speech synthesis [4,5] belongs to the group of the concatenative speech synthesizers. The synthetic speech signal is generated

by concatenation of short (typically pitch size) waveform elements. Modification of the length and size of each sound element and smoothly connecting them can generate high quality speech signal 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 diphonems 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 quality of synthetic speech can be obtained, but this will increase the size of the database and the time needed to collect it. On the other hand, the speech synthesizer used as speech prosthesis for a patient after tracheotomy should have personal features of the original voice of the patient.

Considering the sound elements of pitch period size, a similarity measure is defined. The similarity 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 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 synthesis of different voices, a new particular voice can be obtained by modification of another one in the existing database. This algorithm can be used for choosing the most suitable one for the modification database. A modified database, which includes vowels most similar to a particular voice, could give synthetic speech of better quality.

2 Auditory 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. The beginning and the end of the selected signal were placed at zero-crossing points to obtain a smooth connection in the concatenation. The collected elements were standardized in length and amplitude by multiplying with a trapezoidal window to increase the pitch or by adding a short silence for decreasing the pitch. So all played sounds had the same length and amplitude.

By concatenating each stored element 40 times, 47 different steady-state sounds were generated. 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 one 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 listeners.

Table 1. Results of the auditory test: similarity of sounds pairs determined by the listeners.

Sound	Similarity %	Sound	Similarity %	Sound	Similarity %
1-1	100.0	1-17	4.55	1-33	54.55
1-2	4.55	1-18	4.55	1-34	0.0
1-3	9.09	1-19	22.73	1-35	18.19
1-4	0.0	1-20	45.45	1-36	0.0
1-5	31.82	1-21	9.09	1-37	36.36
1-6	81.82	1-22	18.19	1-38	4.55
1-7	18.19	1-23	16.64	1-39	0.0
1-8	9.09	1-24	16.64	1-40	16.64
1-9	4.55	1-25	0.0	1-41	0.0
1-10	16.64	1-26	0.0	1-42	0.0
1-11	68.18	1-27	31.82	1-43	16.64
1-12	4.55	1-28	36.36	1-44	4.55
1-13	0.0	1-29	36.36	1-45	68.18
1-14	0.0	1-30	4.55	1-46	4.55
1-15	27.27	1-31	16.64	1-47	4.55
1-16	31.82	1-32	9.09		

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 other ones (Table 1). If we look at the histogram of similarity (Fig. 1) 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 decision was placed, while the horizontal axis describes the sound number that was compared with the sound number one.

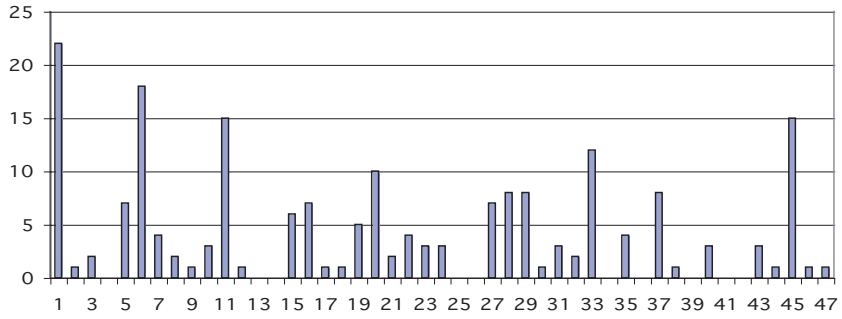

Fig. 1. Auditory test results as a histogram.

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 triphone CVC), the listeners noticed significant differences. The perceptual differences can be computationally estimated by using proper auditory models. We have applied an auditory spectrum algorithm that warps the Fourier spectrum to the Bark scale and the magnitude scale to loudness density (also called specific loudness) [3,7]. Considering the auditory spectra of similar and different pairs we can see large differences in the spectral patterns. Figures 2 and 3 present the auditory spectra of two sound pairs. Fig. 2 shows the auditory spectra of two similar sounds, while Fig. 3 presents auditory spectra of two different sounds.

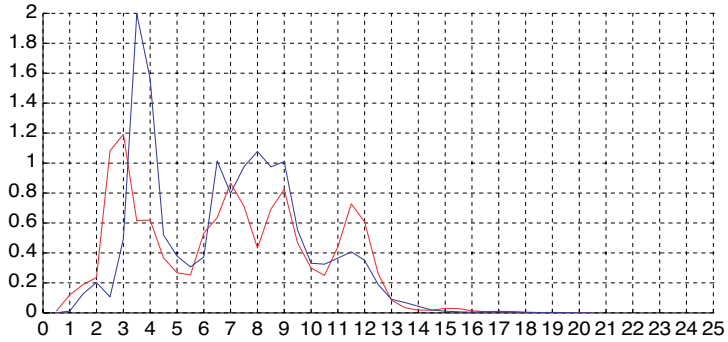

Fig. 2. Auditory spectra of two similar sounds. Horizontal axis: Bark scale; vertical axis: loudness density [sone].

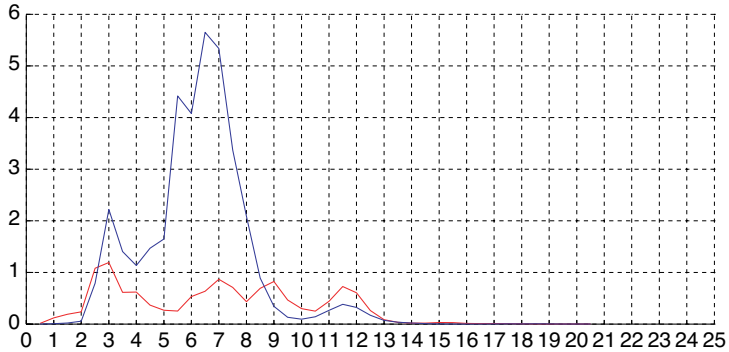

Fig. 3. Auditory spectra of two different sounds; Horizontal axis: Bark scale, vertical axis: loudness density [sone].

4 Similarity Determination Algorithm

For finding the similarity of two elementary sounds the 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 [1] 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 [1]. The width of a minimum region was then set to 1 Bark around the minimum. 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 sounds 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. Additional adjustment was applied to "tune" the algorithm to match the auditory tests result.

5 Parameter Histograms

Each one of the recorded voices comes form 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 the histogram of Fig. 4, 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.

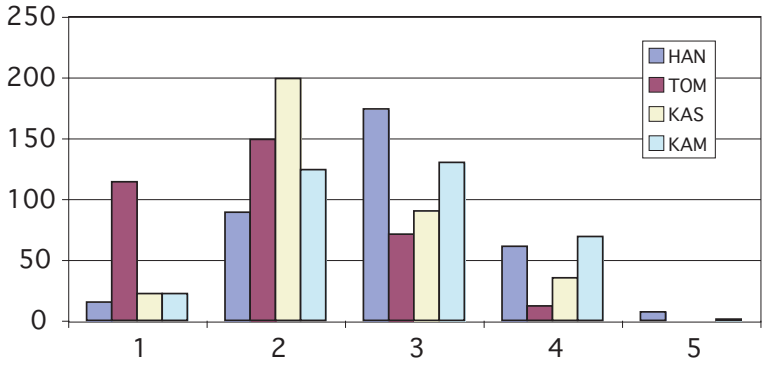

Fig. 4. Histogram of numbers of bands (horizontal axis) in vowel 'a' for four different voices.

Another example is the positions of the centre of gravity. Figure 5 presents the centre of gravity position of the first band for one voice in vowel 'a'.

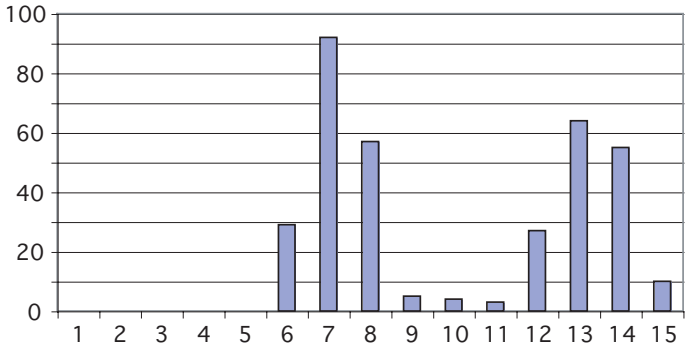

Fig. 5. Histogram of the position of centre of gravity (horizontal axis: Bark scale) of the first band in vowel 'a' for speaker "TOM".

6 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 microphonems for vowel 'a' (from the middle part of the vowel) and 215 microphonems for 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 the single test, the similarity of a short sound pair was calculated. The result of each 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 was taken from the first voice and the second one from the second voice. Thus, 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 in Tables 2 and 3.

Table 2. Similarity of voice pairs for vowel 'a'.

Speaker	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 3. Similarity of voice pairs for vowel 'o'.

Speaker	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 results were 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 received when the vowel 'o' was considered.

7 Discussion

Using this simple method of sound comparison we can notice how many different realizations of each vowel we can meet 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 voice HAN the most common number of energy bands for vowel 'a' was 3 while for voice KAS it was 2. 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 exchange of the vowels and some voiced consonants to obtain another individual sounding voice. In this case the most significant realizations (for a particular speaker) of each vowel have to be applied as the database modification elements.

8 Conclusion

The presented method of analyzing voices requires low computational effort and is based on very short sound signals that are typical for the microphonemic synthesis method. The presented results show that it is possible to compare two voices by analyzing sets of short signals extracted from many different points of a single subject's utterance.

References

1. Chistovich L. A., "Central Auditory Processing of Peripheral Vowel Spectra," J. Acoust. Soc. Am. 77 (3), March 1985, p. 789-805.
2. Karjalainen M., Laine U. and Toivonen R., "Aids for the Handicapped Based on SYNTE 2 Speech Synthesizer," Proc. IEEE ICASSP'80, Denver 1980, p. 851-854.
3. Karjalainen M., "A New Auditory Model For The Evaluation of Sound Quality of Audio Systems," Proc. IEEE ICASSP'85, Tampa 1985, p. 608-611.
4. Lehtinen L. and Karjalainen M., "Individual Sounding Speech Synthesis by Rule Using the Microphonemic Method," Proc. Eurospeech'89, Paris 1989, p. 180-183.
5. Lukaszewicz K. and Karjalainen M., "Microphonemic Method of Speech Synthesis," IEEE, ICASSP'87, Dallas 1987, p. 1426-1429.
6. Wloskowicz D., Lukaszewicz K. and 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.
7. Zwicker E. and Fastl H., Psychoacoustics - Facts and Models. Springer Verlag, 1990.