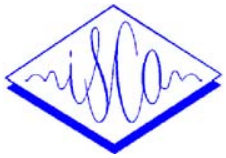

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DISTANCE SCORE EVALUATION OF THE VISUALISED SPEECH SPECTRA AT AUDIO-VISUAL ARTICULATION TRAINING

Vicsi, K. - Csatari, F. - Bakcsi, Zs. - Tantos, A.
Technical University of Budapest, Hungary
vicsi@ttt-202.ttt.bme.hu

10.21437/Eurospeech.1999-419

ABSTRACT

In the frame of the Inco-Copernicus program of the European Commission titled „A Multimedia Multilingual Teaching and Training System for Speech Handicapped children" an audio-visual pronunciation teaching and training method and software system has been developed for hearing and speech-handicapped persons to help them to control their speech production.

During a part of the training the interpretation of the signal is based on the comparisons of the signal with the stored references. The aim of the present study is to find a distance measure that can help these comparisons and mirror the judgement of the listeners. Three spectral distance calculations have been compared. The good and unacceptable examples were separated well on the base of the Average Spectrum Distance calculation. This calculation can be the base of an automatic feedback of the actual pronunciation that could approach the decision of the listeners well.

1. INTRODUCTION

In our articulation-training project developing for hard of hearing children the produced sound is measured and visualised. There is a direct relationship between the articulation and the resulted speech picture (visualised speech parameters) thus, the measured and visualised speech parameters gives information about the position of the articulation organs, about the correctness of the articulation.

During the training the interpretation of the signal is based on the comparisons of the signal with the stored references. The comparison is firstly made visually and evaluated by the human brain, secondly with an automatic feedback, based on the distance calculation between the spectral components of the reference and the actual speech. The aim of the present study is to find such distance measures, that can help these comparisons. The distance scores must be in good agreement with the judgement of most of the people in the child's educational, public and cultural life. The main point is whether people can understand the children easily or with difficulty [1]. We look such a distance score which can mirror the judgement of listeners by investigating the perceptual decisions of different pronunciation qualities varied from the good pronunciation to the defective one. The obtained distance metrics could help in the quantitative evaluation of the defective speech.

2. CORPUS

The speech of 53 children aged from 5 to 10 was recorded. Children pronounced fricative sounds and vowels in one, two and tree syllabic words and in sentences. All examined phonemes occurred with different sound connections, at the start, at the end and at the middle of the words. All together 80 words and 29 sentences were constructed in the text.

The text was recorded in an anechoic chamber, spoken by 5-6 year-old children collected from public nursery schools, and 7-10 year-old children collected from public elementary schools.

Children who could not read repeated the text spoken first by the assistant. We made the sound recordings providing a friendly environment for the kids.

3. PERCEPTION EXPERIMENT

In the corpus, the children were with good and average pronunciation and with impediment in speech. Examples were given in statistical order.

20 speech therapist and 13 non-expert students were asked to classify the examples of the children using 3 categories: good, acceptable and unacceptable.

We examined, at what extent the perceptual decision can be accounted for in terms of the acoustic properties of the stimuli, and which type of the acoustic distance calculation [2], [3], [4], [5] can approximate the results of the perception test. Listeners were asked to concentrate only on one phoneme in a word and classify only this one:

Giving 3 points in case of good pronunciation;
Giving 2 points in case of acceptable pronunciation;
Giving 1 points in case of unacceptable pronunciation;

The average decision of 20 therapist and/or 13 students was calculated according to the following equation:

$$\text{Decision} = \frac{1xN_u+2xN_a+3xN_g}{N_u+N_a+N_g}$$

where N_u, N_a, and N_g are the numbers of the unacceptable, acceptable and good decisions. Three representative groups of phonemes (pronounced in words) were selected from the speech material. Good examples were chosen in those cases where decisions were between 2.8-3, acceptable ones between 1.8-2.2 and unacceptable ones between 1-1,5.

4. FUNCTION OF THE AGE ON THE QUALITY OF THE PRONUNCIATION OF PHONEMES

We evaluated the pronunciation of different phonemes in the function of the age. The listeners show that the pronunciation of fricatives and affricates were not perfect (under acceptable) at the age 5 at normal speaking children and the quality of the pronunciation of these sounds developed until 10. as it is

presented in Fig. 1 and 3. These results can reflect the anatomical and neuro-muscular development of children [6].

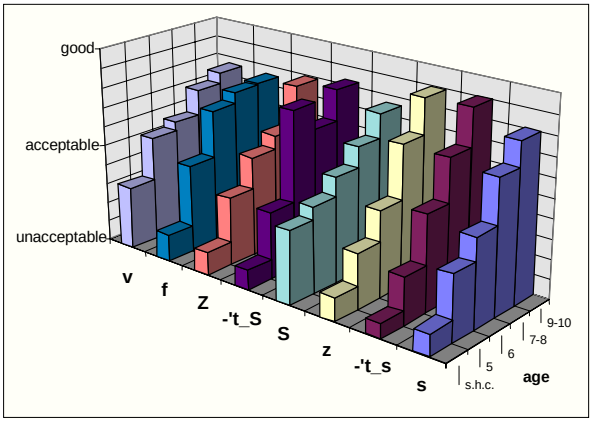

Fig 1. The subjective decisions of each fricative and affricate in the function of age in case of non-expert listeners (shc means speech handicapped children). (Phonemes are marked with SAMPA symbols.)

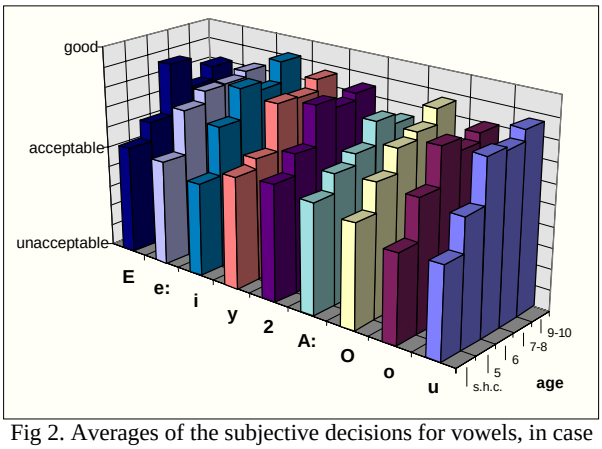

Fig 2. Averages of the subjective decisions for vowels, in case of non-expert listeners (shc means speech handicapped children)

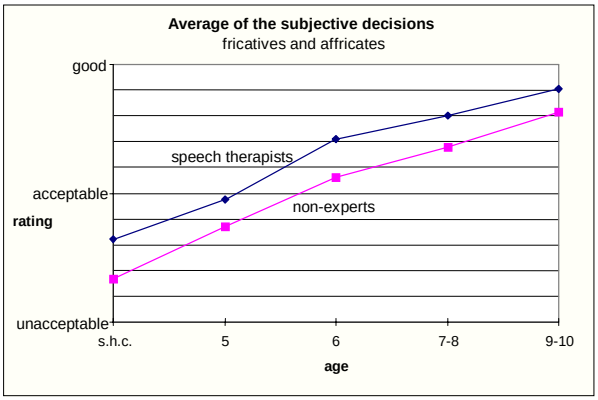

Fig 3. The average of the subjective decisions of all fricatives and affricates in function of the age of the children

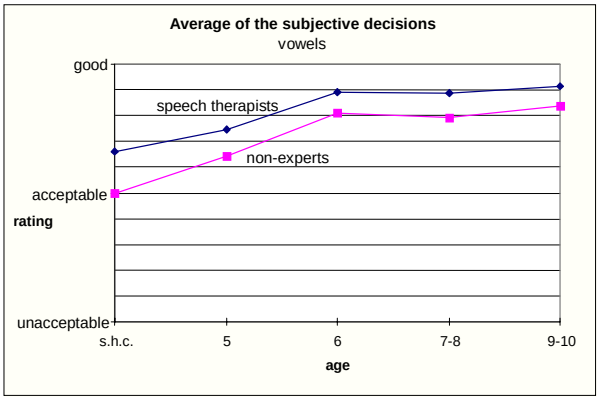

Fig 4. The average of the subjective decisions of vowels in the function of the age of children

The definitive pronunciation of vowels evolves earlier than that of consonants. It was found through the listening test – as it is shown in Fig. 2 and 4. – that the quality of vowels at age 5 was judged to be not quite perfect, but it was found definitely good by the age of 6.

This psycho-acoustical result, which presents the speech development of the children by the decisions of listeners, is correlated well with acoustical cross-sectional study of childrens speech [8]. This study confirms, that the reduction of magnitude and within–subject variability of temporal and spectral parameters with age is general acoustic phenomenon associated with the speech development of children.

It is worth mentioning that speech therapists were more permissive than non-experts. We must take the decisions of non-experts into consideration, because children have to make themselves understood with non-expert persons in the real life.

5. SPECTRAL DISTANCE CALCULATION

In our spectral analysis, the decomposition of the complex speech sounds into their component frequencies has been done in critical filter bands, from 80Hz to 8 KHz, thus the spectral resolution was 1 Bark.

Three different types of the distance computations based on this spectral measurement were compared.

5.1 Average spectrum distance

Here we calculated the distance of the spectrum components of an actual pronunciation from the averaged good examples.

The boundaries of the examined phonemes in the words marked with our automatic segmentation and labeling program LIAS [7], and the average spectrum of each good phonemes were calculated the way that the first and the last two frames of all phonemes were omitted. Thus we have got the average, the spread and extreme values of the spectrum of each phoneme as it is shown in the case of sound ‘i’ in Fig. 5.

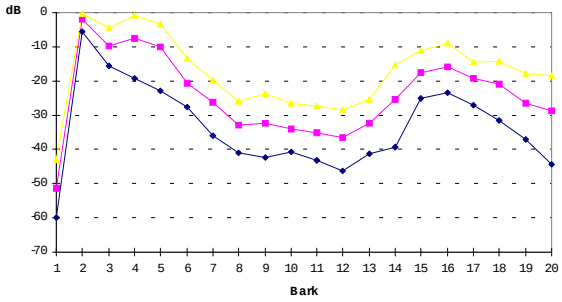

Fig 5. The average and the extreme values ($Spectr_{i\,max}$, $Spectr_{i\,min}$) of the spectrum of the phoneme 'i' (in words, in different positions and sound connections.)

The $D_{specctrum}$ distance of an actual pronunciation ($Spectr_{ia}$), from the extreme values ($Spectr_{i\,max}$, $Spectr_{i\,min}$) of the good examples can be calculated by the following equation:

$$D_{spectrum} = \sum_{i=1}^{20} W_i d_{spectrum\ i} \qquad (1)$$

where W_i are constants, and

$$D_{spectrum\ i} \begin{cases} Spectr_{i\,min} - Spectr_{i\,a}, & Spectr_{i\,a} < Spectr\ i_{min} \\ Spectr_{i\,a} - Spectr_{i\,max}, & Spectr_{i\,a} > Spectr\ i_{max} \\ 0 \text{ otherwise.} & \end{cases} \qquad (2)$$

5.2 Dynamic time warping of spectrograms

The actual distance between the reference example and the examined example was calculated here. We used the well known formula to warp the examined pronunciation to the reference one, and to calculate the minimal distances (D_{din}) between them.

$$D_{din} = \min \begin{cases} g(i, i-1 + d(i, j) \\ g(i, i-1, j-1) + 2d(i, j) \times \dfrac{1}{N_I \times N_J} \\ g(i-1, j + d(i, j) \end{cases} \qquad (3)$$

where N_I and N_J are the frame numbers of the actual and the reference examples.

During one kind of the speech training the speech picture of a nice good reference was compared with the speech picture of the actual one. So not an average picture was presented. Thus by the automatic feedback the same examples are used for the calculation and presented on the computer screen. This is the reason why we wanted to examine the usability of the traditional DTW algorithm.

5.3 Spectral component differences

The reference and the examined example are warped to each other. In the case of the best fitting we calculate only the number of those spectral points at which the difference of the corresponding spectral components of the corresponding frame is bigger than a threshold. We called this distance as the picture point distance ($D_{picture}$), because it tries to approximate the visible difference between the reference picture (spectrogram of the reference sample), and the examined picture (spectrogram of the examined picture).

$$D_{picture} = \frac{M \times 1000}{20\ n_I} \qquad (4)$$

M is the number of spectral points at which the difference $^3d(i\ j)^3>$ threshold in dB.

n_I is the number of frames of the reference example

6. EVALUATION OF THE RESULTS OF THE DISTANCE COMPUTATIONS

The examples of the perception test in the good, acceptable and unacceptable groups of each phonemes were used for the distance calculations.

Distances were calculated within the phonemes and among the phonetically similar phonemes. The obtained distance averages and spreads are presented in Table 1 and 2.

Analysing the distribution of the distances of the different categories in case of the three distance computations, it was found that the Average Spectrum distances can be separated well according to the listeners categories. The DTW and Spectral Component Differences are more worse. Comparing the last two methods the Spectral Component Difference seems to be better than DTW.

The phonemes that were judged good and the phonemes that were judged unacceptable can be separated well on the base of the Average Spectrum Distance ($D_{spectrum}$) and the good examples can also be distanced from the phonetically similar phonemes. The acceptable examples could not be separated well either from the good or from the unacceptable examples, but the listeners themselves used this category when they were uncertain.

The results show that a good automatic feedback of the actual pronunciation can be made on the base of the $D_{spectrum}$ calculation. This feedback can approach well the decision of the listeners.

Table 1. Average Spectrum Distances ($D_{spectrum}$) within the phoneme and among the phonemes

within the phoneme	Examined phonemes (SAMPA)											
	s		z		S		Z		i		e:	
	average	spread	average	spread	average	spread	average	spread	average	spread	average	spread
good	0	0	0	0	0	0	0	0	0	0	0	0
acceptable	151	132	87	47	113	83	532	298	23	13	180	144
unacceptable	430	270	250	242	953	592	980	267	226	244	162	0
among the phonemes												
s	0	0	156	84	712	309	2134	556				
z	276	200	0	0	2027	1114	1344	400				
S	498	120	353	114	0	0	1600	365				
Z	900	158	154	30	2331	374	0	0				
i									0	0	156	119
e:									103	57	0	0
E									140	285	375	117
y									150	63	89	57

Table 2. Dynamic time warping distances (D_{din}), and Spectral component distances ($D_{picture}$) within the phoneme and among the phonemes

within the phoneme	D_{din}				$D_{picture}$			
	s		i		s		i	
	average	spread	average	spread	average	spread	average	spread
good	20	6	20	16	25	7	33	18
acceptable	25	3	50	10	25	11	33	18
unacceptable	25	14	50	12	33	9	50	21
among the phonemes								
s	0	0			0	0		
z	25	18			25	18		
s	25	3			33	12		
z	100	21			50	25		
i			0	0			0	0
e:			33	13			25	11
E			25	24			33	16
y			33	13				16

ACKNOWLEDGEMENTS

The work has been supported by the Hungarian Scientific Research Foundation, and by the European Community as a part of the „SPECO"-Copernicus program (No: 977126).

REFERENCES

[1] J.L. Wallace at al. (1998): "Applications of Speech Recognition in the Primary School Classroom" ESCA - Still 98 Workshop Proceedings, Marholmen, pp. 21-24.

[2] Sakoe, H. and Chiba, (1978), "Dynamic programming algorithm optimalization for spoken word recognition", IEEE Tranz. Acoust. Speech and Signal Process, Vol. ASSP-26, pp. 43- 49.

[3] Zwicker, E., and Terhardt, E. (1980), "Analytical expressions for band rate and critical bandwidth as a function of frequency", J. Soc. Am. Vol. 68. pp. 1523.

[4] Itakura, F, (1975), "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-23, 67-72.

[5] Blomberg, M. at al.(1986), "Auditory Models as Front end in Speech Recognition Systems", Invariance and variability of Speech Processes, Erlbaum, Hillsdale, NJ, 108-114.

[6] Kent, R.D. and Read, C. (1992) The Acoustic Analysis of Speech, Group Inc. San Diego, California, Singular Publishing, pp. 158-164.

[7] Vicsi, K at al. (1998), „LIAS: Language Independent Automatic Segmentation Technique Using Sampa Labeling of Phonemes. First International Conference on Language Resources & Education, Granada Spain, 1998. 1.317

[8] Lee, S. at al (1999), „Acoustics of Children Speech: „Developmental changes of temporal and spectral parameters", JASA, Vol.3. pp. 1455-1468.