Acta Cybernetica **22** (2016) 771–789.

Adapting Dynamic Time Warping to the Speech of the Hearing Impaired[*]

László Czap[†] and Attila K. Varga[†]

Abstract

One service provided by our application 'Speech Assistant System' assisting the teaching of the hearing impaired to speak is the automatic assessment of words and sentences in the course of practice and feedback to the person. Individual speech sounds can only be correctly evaluated if they are compared with the appropriate reference speech sounds. This requires segmenting the speech to be examined. The methods currently known do not give sufficiently correct results for the speech of the hearing impaired, which is often so distorted and halting so that it prevents understanding. The paper presents a reference generation method suitable for segmenting distorted speech, a modification of dynamic time warping and its comparison with traditional methods. The procedure presented has been successfully used for the automatic assessment of the pronunciation of the hearing impaired.

Keywords: dynamic time warping, speech quality assessment, acoustic-phonetic features, distorted speech, speech of hard of hearing children

1 Introduction

The project 'Basic and Applied Research for the Internet-Based Speech Development of the Hearing Impaired and for the Objective Measurement of Progress' served the purpose of creating a new aid for the deaf and the hard of hearing in learning to speak, called the Speech Assistant System. The foundation of the research is represented by the 'talking head' developed at the University of Miskolc and the audio-visual transcoder developed at the University of Debrecen. The objective of the project is to create a complex system which provides the audio-visual representation of the speech process, by the visual representation of the sound images of speech on the one hand, and of the articulation on the other, set in a

[*] The research has been carried out within the framework of Mechatronics and Logistics Centre of Excellence operating as one of the strategic research areas of the University of Miskolc and as part of the TMOP-4.2.2.C-11/1/KONV-2012-0002 project funded by the European Union, co-financed by the European Social Fund.

[†] University of Miskolc, E-mail: {czap,varga.attila}@uni-miskolc.hu

DOI: 10.14232/actacyb.22.4.2016.3

training framework system. The 3D head model with its transparent face can visualise the tongue motion better than a natural speaker. In addition, the system includes a number of functions (visualisation of prosody, automatic assessment and implementation of the knowledge-based system) that facilitate individual practice not only on the computer, but also on a mobile device. The module of the technology developed performing the audio-visual transcoding is language independent, so the talking head and the automatic assessment can be adapted to other languages besides Hungarian.

Automatic assessment provides feedback to the hearing impaired, who can use the Speech Assistant System on their own. The assessment of the speech produced during practice shows not only progress achieved in utterance, but also serves as input to the knowledge-based system, which assists in designating the next word to be practiced based on teacher experience [1].

The methods developed for automatic speech quality assessment include speech segmentation in explicit or implicit form. Speech tempo changes from speaker to speaker, from articulation to articulation. These non-linear extensions and shortenings do not necessarily count as faulty pronunciation. The hearing impaired usually speak more slowly than the average speech tempo. For the assessment of the pronunciation of the individual speech sounds, the time segments of the reference pattern and of the actual pronunciation have to be matched. The reference and the actual waveforms can be made to have identical lengths by linear stretching and/or linear shrinking. This, however, does not ensure a time parallel of the individual speech sounds, for the pronunciation rhythm may differ from the reference. If certain speech sounds are pronounced longer and others are pronounced shorter, in linear time warping it will not be the speech sound segments that matched with which it should be similar, therefore the comparison will produce false results. The articulation of certain speech sounds differing in time from their ordinary articulation is particularly characteristic of the speech of the hearing impaired. Therefore, for the purpose of comparing the reference and the speech being examined, non-linear time-warping that is needed, procedures and algorithms developed in computer-based speech processing are available for this purpose. These methods work well for high-quality speech and pronunciation acceptable in everyday communication. However, they produce poor results for distorted speech sounds and unusually drawling and halting speech. The paper discusses our segmentation method suitable for low-quality speech that pairs the test and artificially generated reference shape.

2 Non-linear time warping

We can speak of an ideal time comparison if two samples are aligned along the individual speech sounds. This generally accepted method is used in computer-based speech recognition [15].

The hidden Markov model (HMM), by virtue of its characteristics, is suitable for handling the time structure in speech recognition, for the states belonging to the speech sounds pronounced longer simply return into themselves repeatedly. Maier and others [9] [10] have successfully applied the method with adults whose larynx had been removed due to throat cancer and with children born with a cleft lip and palate. In these patients a close relation can be achieved between the subjective and automatic assessments. This was used as a basis for developing PEAKS (Program for Evaluation and Analysis of all Kinds of Speech Disorders), a recording and analysing system for the automatic or manual assessment of utterance disorders and speech impediments.

Typical, easy-to detect utterance disorders are associated with the different disorders. For the speech disorders, the researchers had training models available in sufficient quantities, so it was possible to develop the statistical model necessary for automatic assessment. However, the mispronunciations of the hearing impaired cannot be typified [7]. Our general-purpose speech recognition device based on the HTK Speech Recognition Toolkit [6], which is adapted to 3,600 words and 1,800 sentences recorded with the voices of 60 school children (12-14 year-old primary school children from three special institutions for the teaching of the deaf and hard-of-hearing) proved to be unsuitable for automatic assessment.

Dynamic time warping (Dynamic Time Warping, DTW) was used in the early era of speech recognition for the comparison of the patterns to be recognised and the reference samples. The procedure examines optimum time alignment as the search for the path with minimum length or weight in a given graph. Let us suppose that the x words to be examined consist of k pieces of segments and the data characterising the i-th $(i = 1, 2, \ldots, k)$ segment are summed up in vector x_i . Next these elementary vectors are collected into a matrix in the classification algorithm. Thus the incoming word is characterised by the vector series $x_1, x_2, \ldots, x_k$. Let the vector series $y_1, y_2, \ldots, y_r$ characterise in a similar way the vocabulary element y, with which the incoming word is to be compared. The objective is to produce a vector series $x_{1},_{2}, \ldots,_{r}$ (length r) from the vector series $x_1, x_2, \ldots, x_k$ by repeating some and omitting others, for which the 'distance'

$$D = \sum_{i=1}^{r} d(x_i, y_i) \tag{1}$$

takes its minimum. Here $d(x, y)$ is an arbitrary given distance function. In producing the vector series $x_{1,2,\ldots,r}$, secondary conditions are set, of which the following is a possible version:

- any vector x_i can only be repeated once (thus we can at most double, but not triple the number of vectors);
- if x_i is omitted, its neighbours (x_{i-1} and x_{i+1}) cannot be omitted, thus two neighbouring segments cannot be omitted;
- the order of segments cannot be reversed [4].

The characteristic vectors used as the inputs to the algorithm are provided by feature extraction of speech. The result of segmentation was examined on the basis of feature extraction by means of the usual procedures:

- MFCC: Mel-Frequency Cepstral Coefficients [2],
- PLP: Perceptual Linear Prediction [5],
- MEL band energy: logarithmic energy [13].

In our experiment the references were provided by recorded speech samples of university students of liberal arts participating in competitions of proper pronunciation and school children with proper pronunciation in the same age group as the hearing impaired involved in the experiment. These references were regarded as standard recorded speech samples. The recorded speech samples of the hearing-impaired children were provided by the speech sound database recorded for the purpose of creating the assessment scale.

The database includes 2,421 words (some words occur several times, but with different speakers, therefore their time structures are also different), which were assessed by 13 teachers and 23 students. Every teacher assessed only the recorded speech samples of the pupils of a school different from his own so as to avoid bias resulting from recognising the speaker. The assessors could listen to a recorded speech sample several times and could make comments on the samples. The results were recorded via an Internet application. In the case of the teachers, the basis of the assessment was given by a five-grade scale set worked out by them.

Interpretation of the scale:

- *Unintelligible (1)*: articulation is completely distorted; the vowels and consonants are unrecognisable; the reproduction of the syllable number is not adequate or discernible; breathing and management of breath is faulty; tempo and rhythm are incorrect; the utterance is unmelodious, non-dynamic or too tense.
- *Difficult to understand (2)*: grave distortions, omission of speech sounds, speech sound replacement; only some of the vowels can be discerned; distortions due to insufficient breathing, e.g. too breathy or choked; characterised by irregular, disturbing tonality, rhythm and tempo.
- *Moderately intelligible (3)*: the articulation of vowels is correct, the number of syllables is appropriate; serious speech defects may occur, e.g. dyslalia (the speech impediment in which certain vowels are incompletely formed), nasality, head voice, prosodic inadequacies.
- *Easy to understand (4)*: slight speech defects; slight prosodic inadequacies.
- *Understandable at the same level as the speech of the hearing (5)*: at most 1-2 speech sound defects may occur.

The 23 students had to score the recorded speech samples on a scale of 1-5 on the basis of everyday usage. Three hundred words were chosen out of the 2,421 words for the detailed segmentation analyses. The stock of words chosen is sufficiently varied not only according to the lengths of the words, but also from the aspect of the occurrence of speech sound juncture features, which is characteristic also of the complete word database. The stock of words of the recorded speech samples was prepared by teachers of the hearing impaired, taking the active vocabulary of the individual students carefully into consideration. The 300 words were manually segmented by a speech processing expert, providing the basis for the comparison.

Comparing the result of the segmentation time warped to correctly articulated reference speech with the time data given by the expert provided values that could not be used for poor quality speech. Often completely different results were obtained for the standard reference samples originating from the various subjects articulating the given word.

The cause of the failure was attributed to the deficiencies in dynamic time warping. The application of dynamic time warping for the purpose of speech recognition was neglected because the comparison has to be performed for every conceivable vocabulary element, which is extremely time-consuming. In addition, more advanced decision-making methods compare the speech section to be recognised not with the voice of a given speaker, but compare the element to be recognised with the data of a population of speakers using statistical learning methods. Our solution integrates a statistical model into the input data, and eliminates the speaker dependence of the reference sample by a new method of reference generation.

It was supposed that the characteristics obtained for the individual speech sounds by statistical methods would provide more reliable results. A neural network was trained on the basis of the BABEL speech sound database.

The BABEL database consists of three different parts: recorded speech samples of numbers of isolated and connected words, CVC (consonant-vowel-consonant) syllables, and continuously read speech. Both the sentences read and the number series were planned so as to provide a good coverage of the speech sound combinations in the Hungarian language. Some of the speech samples in the continuous part are in whispers. Part of the database is segmented into phonemes and labelled. The database includes the voices of a total of 30 male and 30 female speakers as well as 2,000 sentences and 14,000 connected number series.

3 Dynamic time warping input data

We attempted to derive the essence of speech sounds by means of the output activity of neural networks. Neural networks were trained in acoustics-phonetics classification, then using their outputs, new neural networks were trained to differentiate within the class. In the course of training the correct outputs were given a value 1 in their own time frame, and the others were given the value 0. The goodness of classification was checked on testing patterns not included in the training and amounting to a quarter of the complete speech sound material. In the course of testing, in order to obtain the goodness criterion for each speech sound, the sum of the activities of their own outputs was divided by the sum of the activities of the other outputs, calculated for all the testing time segments.

$$G_i = \sum_{\forall R} O_{NN} \Big/ \sum_{\forall F} O_{NN}, \tag{2}$$

where

G_i – goodness of neural network for feature of speech sound or class of speech sounds (i),

O_{NN} – neural network outputs,

$\forall R$ – correct output activity for all of its own time frames,

$\forall F$ – incorrect output activity for all the other time frames.

The neural network whose goodness factor was the maximum for all speech sounds was kept, so we had five neural networks for acoustic-phonetic classification and four neural networks for grouping within the class. For orthography transcription we will use SAMPA symbols. The classes formed by the neural networks are as follows:

- pause;
- vowels *(a, a:, E, e:, i, o, 2, u, y)*;
- semi-vowels *(m, n, J, r, l, j)*;
- fricatives *(f, s, S, h, v, z, Z)*;
- plosives *(p, t, ts, tS, t', k, b, d, d', g).*

The speech sounds belonging to the outputs of the neural network dedicated to the classes are listed in parentheses. We tried to perform the acoustic-phonetic classification by using only a single neural network, but we got weaker results than when using neural networks dedicated to individual classes (Figure 1). For dynamic time warping, these outputs were directly used as a feature vector of the word analysed. Among the speech feature extraction methods examined (MFCC, PLP, MEL subband energy), PLP showed the highest goodness factors, thus the outputs of neural networks trained by PLP speech feature extraction were used as the inputs of dynamic time warping. Training was performed with several options. The setting providing the maximum of the goodness factor is: to the 12 PLP data and logarithmic energy of the actual 40ms frame were added the average of two frames of the preceding 80 ms section and the average of two frames of the subsequent 80 ms. The feature 3 × 13 describes the 40ms segment in the middle of the 200 ms interval. Training of the 5 neural networks meant for phonetic classification was performed with these parameters.

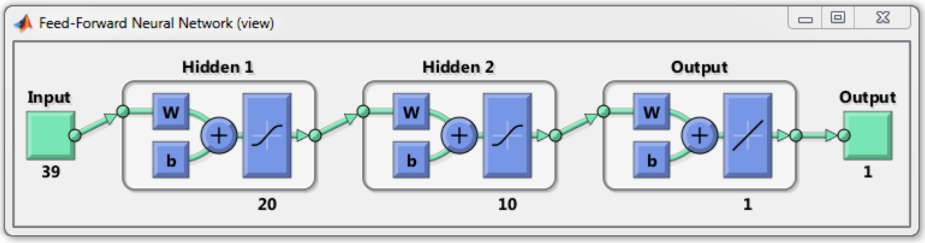

Figure 1: Model of the neural network determining the acoustic speech sound class

In addition to the 39 PLP features, neural networks trained to recognize speech sounds within the phonetic classes were also given the outputs of the five classifying neural networks as input. Figure 2 shows the structure of the neural network used for sorting vowels.

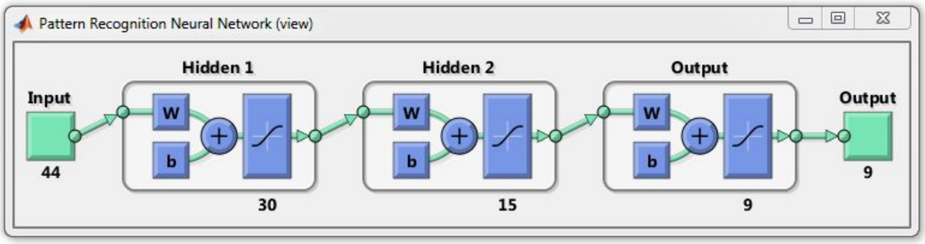

Figure 2: Neural network model of the acoustic speech sound class of vowels

Segmentation was also performed with the neural networks trained with the shorter PLP time frames; the smallest errors were obtained with the above setup. The relevant toolboxes of the program package MATLAB were used for the calculations [11].

4 Reference generation

Using a concrete recorded speech sample as reference, the failure of segmentation discussed above was attributed to individual differences. On the basis of a statistics model, a neural network trained with a great number of speakers is better at reflecting the similarities to the individual speech sounds.

In developing the reference shape, the chosen input data had to be accommodated. Since in this task the objective is not recognition of the word but the alignment of the recorded speech sample, the phonetic transcription of the word was at our disposal. For the purpose of reference generation, the output belonging to the given speech sound and output of the class including the speech sound are made active in the allocated time interval. The timing of the individual speech sounds can be determined starting out from the average time length of the speech sounds [12]. The speech of hearing-impaired children is slower than the average speech tempo. According to our measurements, the time length of the fastest speech was one and a half times the average, therefore in reference generation one and a half times the average time lengths of speech sounds were used, thus a few speech sounds pronounced shorter also fall in the range allowed. Calculating with the above auxiliary conditions of dynamic time warping, the length of individual speech sounds may vary between 3/4 of and three times the average speech sound time length after time warping. Figure 3 shows the created reference features of the word hűséges [hy:Se:gES] (meaning 'faithful'). The horizontal axis shows the time index of the frames, and the vertical axis shows (from bottom to top) the acoustic-phonetic features starting with the pause then the outputs classifying vowels, semi-vowels, fricatives and plosives, and above them the individual outputs of the speech sound classes in the above order.

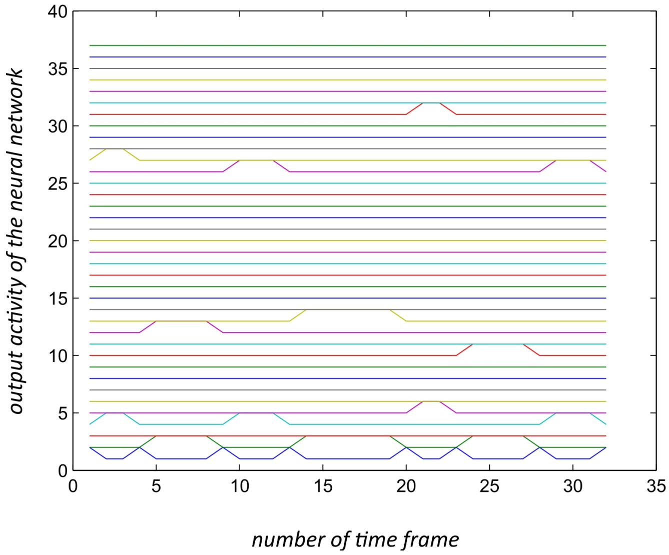

Figure 3: Features created in the reference generation of the word [hy:Se:gES]

In case of good quality speech, the outputs of the neural networks show significant activity. The word 'hűséges' can be clearly understood and is of a quality accepted in everyday communication (Figure 4).

On the other hand, the output levels decrease visibly and several outputs show activity simultaneously as Figure 5 shows the output activities belonging to the word 'valami' [vOlOmi] (meaning 'something') pronounced with a distortion making it unintelligible. The outputs of a neural network are not faultless and among the speech sound samples to be segmented there are also speech samples distorted to unintelligibility.

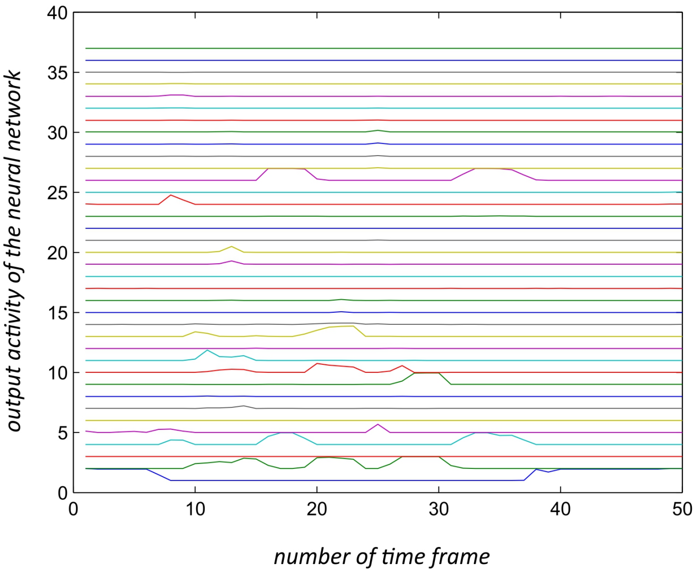

Figure 4: Significant activity of the actual outputs
in aligning the clearly understandable word [hy:Se:gES]

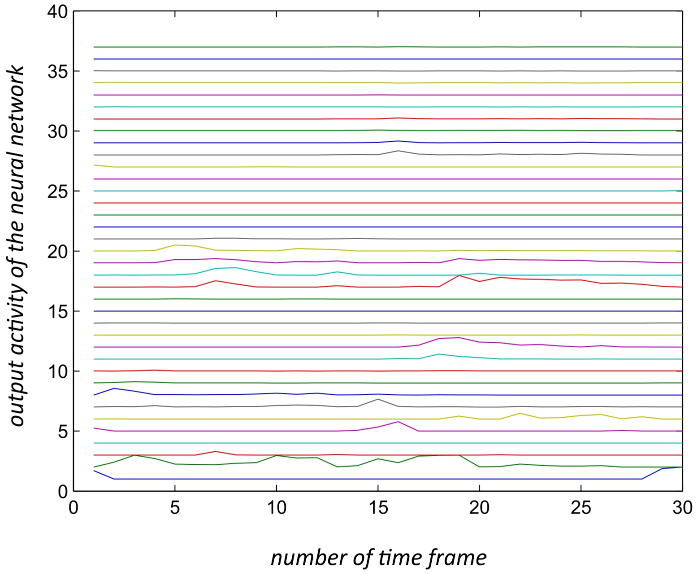

Figure 5: Weak activity of the actual outputs
in aligning the word [vOlOmi] with a distorted pronunciation

Therefore not simply the activity of the relevant output is used as input, but the activities of the outputs belonging to the same acoustic-phonetic class are summed and weighted with the similarity measurement between the speech sounds.

On the basis of the speech sound database BABEL and using PLP speech feature extraction, the average of the coefficients belonging to the total occurrence of the individual speech sounds was determined. A 40 ms segment was marked from the center of each speech sound having a stationary phase (vowels, semi-vowels and fricatives) and the last 40 ms (burst) for plosives. Then Euclidean distances were formed between the averages of Hungarian speech sounds [3]. By reversing the normalised distance, similarity measurements were formed between the individual speech sounds:

$$H(i,j) = 1 - D(i,j)/D_{max}, \tag{3}$$

where $H(i,j)$ is the similarity of the i-th and j-th speech sounds, $D(i,j)$ is the distance of the averages of the PLP coefficients of the i-th and j-th speech sounds and D_{max} is the maximum of these distances.

Summation of the similarity measurements for an acoustic-phonetic class is as follows:

$$S(i) = \sum_{j \ni O} H(i,j) * NN(j), \tag{4}$$

where O designates the speech sounds belonging to its own class and $NN(j)$ is the j-th output of the neural network.

Without transferring output activities to the similar phones, in case of misclassification or highly distorted speech the right neural network output would not get any activity, causing false pause frames in the time interval of the phone. This time shift would risk the right segmentation of neighbouring phones as well. The artificially created reference pattern and the cumulated neural network outputs of speech examined form the basis of dynamic time warping.

The following figures show the similarity measurements of speech sounds belonging to the classes of the neural network compared to each other (Figures 6–9).

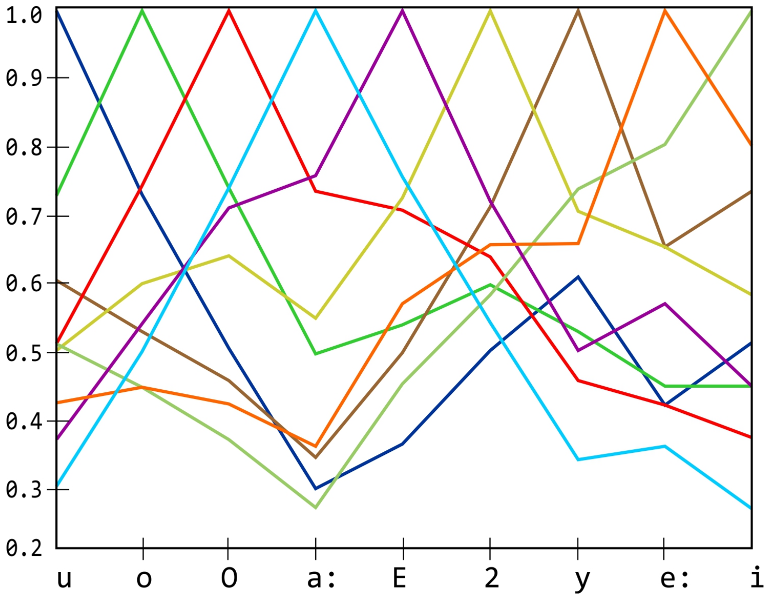

Figure 6: Similarity measurements of vowels

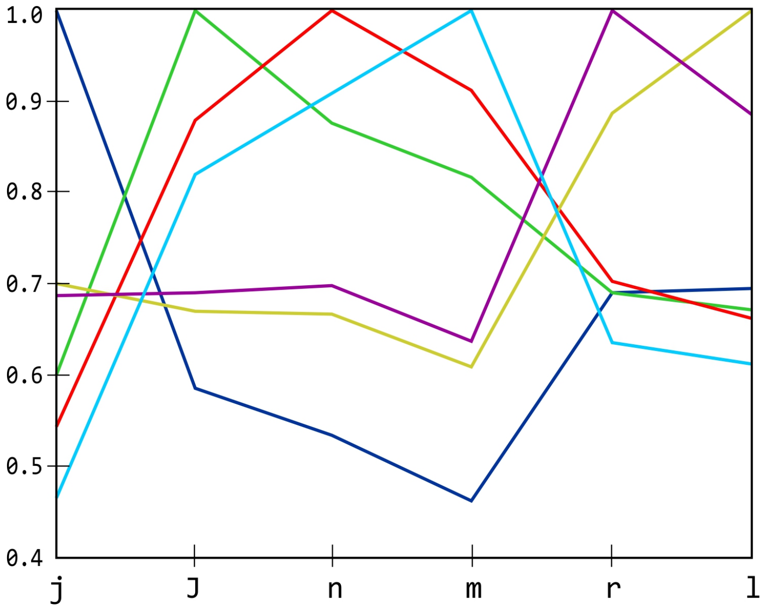

Figure 7: Similarity measurements of semi-vowels

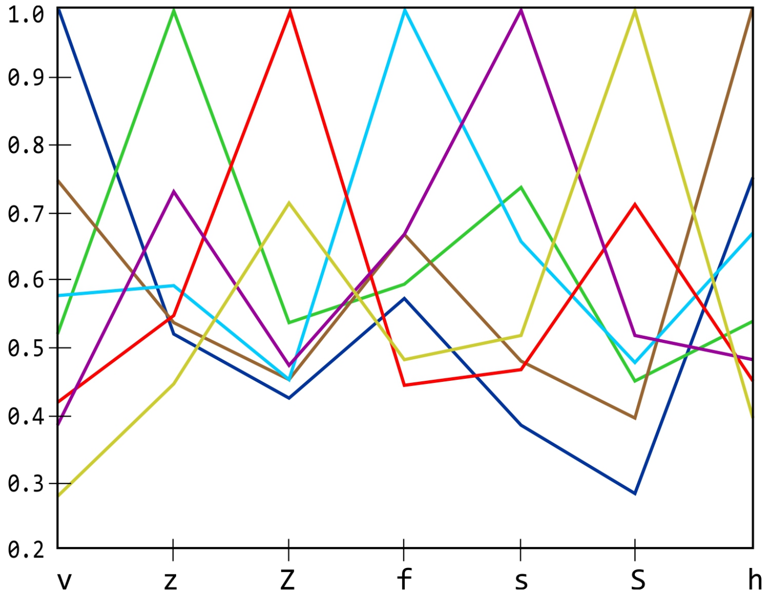

Figure 8: Similarity measurements of fricatives

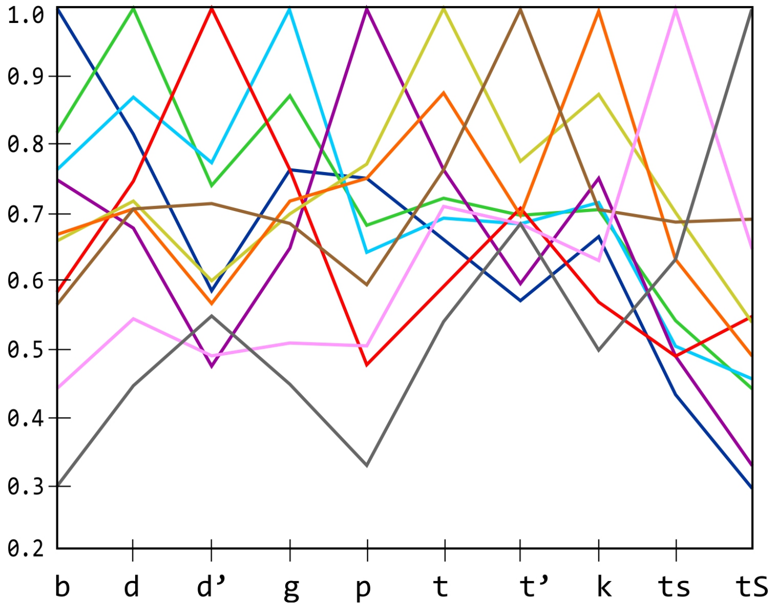

Figure 9: Similarity measurements of plosives

5 Modifying the DTW algorithm

Tested by the classical rules of dynamic time warping and using the outputs of the neural network as feature vector, time warping produced much better results than when recorded words were used as reference.

During testing, however, it was found that in the speech samples of hearing-impaired children pauses of several tenths of a second frequently occurred between speech sounds. In order to treat this problem, the secondary conditions of the dynamic time warping algorithm were modified:

- an optional pause was inserted after each speech sound in producing the reference;

- the pause can be repeated number of times.

According to the rules set out above, a time interval can be lengthened to a maximum of twice its original length. However, in the speech samples of hearing-impaired children there were often speech sounds pronounced longer than that.

Therefore:

- double reiteration of a time frame can also be allowed, thus a time interval can be lengthened to three times its original length. In the following, this will be referred to as adapted dynamic time warping (ADTW).

Figure 10 shows the segmentation results of the haltingly pronounced word ’lázmérő’ [la:zme:r2:], (meaning ’clinical thermometer’) and the hardly intelligible word ’valami’ [vOlOmi], (meaning ’something’) as examples.

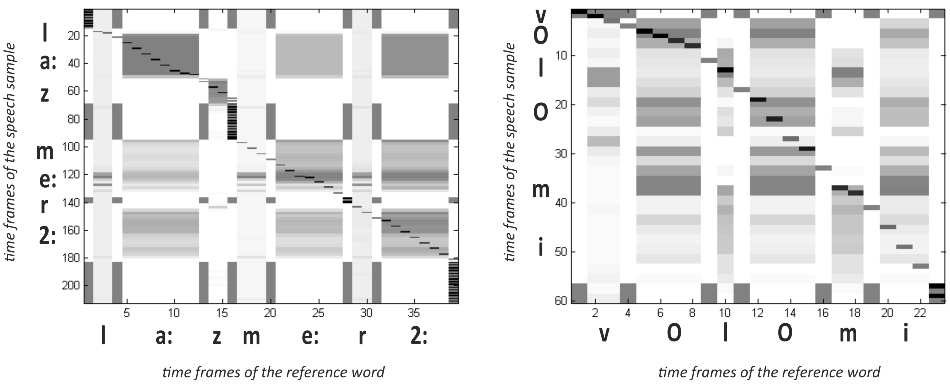

Figure 10: Dynamic time warping of
the words [la:zme:r2:] and [vOlOmi] using the ADTW method

The horizontal axis shows the reference segments and the vertical axis shows the segments of the speech sample examined. In the reference the vertical bands of the size of one time frame indicate the pauses inserted. The shading of the points of the matrix is proportional to the similarity measurement. Darker bands mean greater similarity. The horizontal line appears in the time frames matched. On the horizontal axis of the word 'lázmérő' [la:zme:r2:], time frame 16 shows the insertion of a pause of considerable length. The word 'valami' [vOlOmi] even with an extremely distorted articulation can be successfully segmented with the procedure proposed.

The recorded speech samples can be heard at the following links:

`http://mazsola.iit.uni-miskolc.hu/~avarga/hangmintak/huseges.wav`
`http://mazsola.iit.uni-miskolc.hu/~avarga/hangmintak/lazmero.wav`
`http://mazsola.iit.uni-miskolc.hu/~avarga/hangmintak/valami.wav`

6 Evaluation

In evaluating the modified algorithm, expert segmentation was regarded as the reference. In expert segmentation, the segment boundaries were determined on the basis of a combination of the time function, the spectrogram and the speech sound played from the segment boundary (or to the segment boundary). The comparison was performed using two other segmentation procedures:

1. DTW algorithm based on acoustic-phonetic features, optimised for good-quality speech without being adapted to hearing-impaired speech samples.

   The objective of the time warping method based on acoustic-phonetic (AF) speech sound classes is to compare prosody (the combination of melody, pronunciation speed, rhythm, stress, speech sound intensity and tonality), which can be applied to several languages. It is a time warping method which aligns the two samples strictly along the speech sounds and performs scaling only within them. In this method the novelty is represented by the execution method of the computer segmentation, for which general acoustic speech sound classes were used, which determined language-independent articulation configurations [8]. Applicability to several languages followed from that. In the present case the developers supposed that the difference between the actual and the reference sample is minimal (the speaker is cooperative). Three differences were taken into account: insertion, omission and different pronunciation. The AF segmentation procedure was not adapted to poor-quality speech.

2. HMM-based speech recognition with PLP feature extraction, with the pauses between the speech sounds and their repetition of optional times included in the grammar rules and forced alignment segmentation.

Table 1: Proportions of correct segmentation for different speech feature extraction methods

Tolerance (ms)	**Speech feature extraction methods**		
	MFCC	**PLP**	**MEL**
<= 10	0.31	0.48	0.38
<= 20	0.61	0.72	0.60
<= 30	0.78	0.84	0.73
<= 40	0.86	0.90	0.82
<= 50	0.90	0.93	0.88
<= 60	0.93	0.94	0.91
<= 70	0.95	0.95	0.93
<= 80	0.95	0.95	0.93
<= 90	0.96	0.95	0.94
<= 100	0.96	0.96	0.94

The recordings of the 24 male and 24 female speakers of the BABEL speech sound database provided the training samples, and recordings of 6 male and 6 female speakers provided the testing samples. A 10 ms time frame was chosen and the previously used speech feature extraction procedures were examined as feature vectors from a segmentation aspect:

- MFCC: 12 coefficients and log energy give the 13 components,
- PLP: 12 coefficients and log energy,
- MEL: logarithmic band energy dividing the 125 Hz – 8 kHz frequency domain into 30 part bands on the basis of the mel-scale.

On the basis of the results (Table 1), PLP speech feature extraction was chosen here and will be referred to as HMM in the following. Since the ultimate objective of the method to be developed is automatic assessment, in speech feature extraction the centre of phones is searched for, thus the stationary phase if there is one will characterise the given speech sound [14]. Therefore a segmentation error is regarded as serious or less serious depending on its sign. If the segmenter puts the beginning of a speech sound further forward of the real limit, the error is in the incorrect direction, for the erroneous limit lies outside of the interval of the desired speech sound. Again the error is more serious if the end of the speech sound is put farther back of the real limit. The error is not so serious if the limit is placed farther back of the real beginning or further forward of the real ending within the speech sound examined (Figure 11).

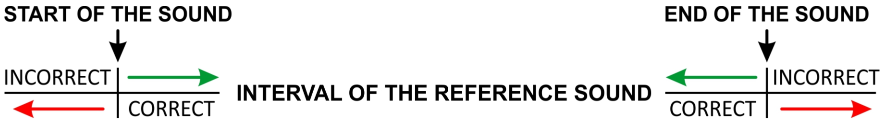

Figure 11: Segmentation error by error direction

The following tables (Table 2–5) sum up the results of the segmentations performed by the different procedures concerning the 3,694 speech sounds included in the 300 words.

Table 2: Results of the acoustic-phonetic (AF) segmentation procedure

AF segmentation procedure				
Tolerance (ms)	**Initial**		**Final**	
	Incorrect	**Correct**	**Correct**	**Incorrect**
0	1785	62	1782	65
20	1747	100	1795	52
40	1667	180	1801	46
60	1500	347	1805	42
80	1340	507	1811	36
100	1187	660	1812	35
200	692	1155	1825	22

Table 3: Results of the HMM segmentation procedure

HMM segmentation procedure				
Tolerance (ms)	**Initial**		**Final**	
	Incorrect	**Correct**	**Correct**	**Incorrect**
0	1327	528	1569	286
20	620	1235	1669	186
40	296	1559	1734	121
60	190	1665	1761	94
80	142	1713	1780	75
100	122	1733	1792	63
200	66	1789	1824	31

Table 4: Results of the ADTW segmentation procedure

ADTW segmentation procedure				
Tolerance (ms)	**Initial**		**Final**	
	Incorrect	**Correct**	**Correct**	**Incorrect**
0	137	1718	1717	138
20	97	1758	1753	102
40	80	1775	1776	79
60	64	1791	1791	64
80	54	1801	1803	52
100	46	1809	1815	40
200	25	1830	1838	17

Table 5: Number of errors outside of the time interval of
the speech sound from the shifts in the correct direction in the tables above

Segmentation Procedure	**Initial Shifted**	**Final Shifted**
AF segmentation procedure	15	1258
HMM segmentation procedure	135	42
ADTW segmentation procedure	39	77

The results show that 'incorrect direction' errors definitely lying outside of the time interval of the speech sound are a magnitude smaller for the proposed ADTW segmentation than for the AF procedure not adapted to poor quality speech or for the HMM procedure. 'Correct direction' errors, exceeding the time length of the speech sounds are also the fewest also with the ADTW method.

The AF segmentation considered the speech sounds shorter than their real length: placing the beginning of a speech sound is shown in the column of errors in the incorrect direction of Table 3, placing the end of a speech sound more forward is located in the field outside of the domain of Table 5. In HMM segmentation mainly the accuracy of marking the limits at the beginning of speech sounds lags behind the results of the ADTW procedure.

ADTW segmentation is utilized in automatic speech assessment. Speech samples of hearing impaired children were evaluated by the 36 assessors. The average scores served as a reference. The scores of the automatic assessment were closer to the averages than that of 28 subjects out of the 36 ones, while one teacher and seven students reached scores closer to the averages [13].

7 Summary

Adaptation of dynamic time-warping has been presented with the objective of a more efficient segmentation of the voices of hearing-impaired children. Pauses inserted between the speech sounds which can be repeated arbitrarily are able to handle the long pauses of halting speech. Time frames that can be repeated twice – that can be included a maximum of three times – make it possible to follow extremely slow speech. The acoustic-phonetic features used as inputs of the algorithm are able to create perceptible activity at the outputs of the neural networks, thus a statistical model is incorporated into the input data. The proposed method of reference generation does not require the recording of reference speech samples, thus eliminating the speaker-dependence of the reference sample.

References

[1] Czap, L. and Pintér, J. Segmentation of Poor Quality Speech. (in Hungarian). *Proceedings of XX-th International Sciencific Conference of Young Engineers*, pages 119–122, Kolozsvár, 2015.

[2] Davis, S. B., and Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4): 357–366, 1980.

[3] Deza, E. and Deza, M. *Dictionary of Distances.* Elsevier Science Publishers, Netherlands, ISBN 0444520872, 2006.

[4] Faragó, A., Fülöp, T., Gordos, G., Magyar, G., Osváth, L. and Takács, Gy. *Simple isolated word speech recognizer.* (in Hungarian) Research report, 1985.

[5] Hermansky, H. Perceptual linear predictive (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87(4): 1738–1752, 1990.

[6] Hidden Markov Model Toolkit (HTK) website. http://htk.eng.cam.ac.uk/

[7] Illésné K. M. Positive and negativ feedback of Internet-based speech development of the hearing impaired. (in Hungarian). *Alkalmazott Nyelvészeti Közlemények*, 9(1): 135–143, 2014.

[8] Kiss, G., Sztahó, D. and Vicsi, K. Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features. *Proceedings of CogInfoCom2013*, Budapest, 2013.

[9] Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M. and Nöth, E. PEAKS – A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5): 425–437, 2009.

[10] Maier, A., Hönig, F., Hacker, C., Schuster, M. and Nöth, E. Automatic evaluation of characteristic speech disorders in children with cleft lip and palate. In *Proceedings of 11th International Conference on Spoken Language Processing*, Brisbane, Australia, pages 1757–1760, 2008.

[11] MathWorks website. http://www.mathworks.com

[12] Németh, G. and Olaszy, G., editors. *Hungarian Speech.* (in Hungarian), Akadémiai Kiadó, Budapest, ISBN 978 963 05 8755 6, 2010.

[13] O'Shaughnessy, D. *Speech communication: Human and Machine.* Addison-Wesley, U.S.A., ISBN 0201165201, 1987.

[14] Pintér, J. M. *Automatic Assessment of Speech Quality.* PhD thesis (in Hungarian), University of Miskolc, Miskolc, 2015.

[15] Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26 (1): 43–49, 1978.

Received 25th November 2015