Comparison of HMM and DTW methods in automatic recognition of pathological phoneme pronunciation

Robert Wielgat [1], *Tomasz P. Zieliński* [2], *Paweł Świętojański* [1], *Piotr Żołądź* [1], *Daniel Król* [1], *Tomasz Woźniak* [3], *Stanisław Grabias* [3]

[1] Department of Technology, Higher State Vocational School in Tarnów, Tarnów, Poland
[2] Department of Telecommunications,
AGH University of Science and Technology, Krakow, Poland
[3] Division of Logopedics and Applied Linguistics,
Maria Curie-Skłodowska University, Lublin, Poland

rwielgat@poczta.onet.pl, tzielin@agh.edu.pl, p.swietojanski@gmail.com,
pzoladz@gmail.com, danielkrol@poczta.onet.pl, twozniak@vp.pl, grabias1@vp.pl

Abstract

In the paper recently proposed Human Factor Cepstral Coefficients (HFCC) are used to automatic recognition of pathological phoneme pronunciation in speech of impaired children and efficiency of this approach is compared to application of the standard Mel-Frequency Cepstral Coefficients (MFCC) as a feature vector. Both dynamic time warping (DTW), working on whole words or embedded phoneme patterns, and hidden Markov models (HMM) are used as classifiers in the presented research. Obtained results demonstrate superiority of combining HFCC features and modified phoneme-based DTW classifier.

Index Terms: Human factor cepstral coefficients, Mel-frequency cepstral coefficients, dynamic time warping, hidden Markov models, logopedic therapy

1. Introduction

In case of pathological pronunciation of sounds (paradigmatic disorders of speech) the linguistic classification of speech disorders [1] is accepted. Paradigmatic disorders are divided into: (1) elision (no realization of phoneme) (2) substitution (realization of phoneme replaced by realization of other phonemes) (3) deformation. The speech recognition task in this case concerns recognition of selected phonemes embedded in the utterance.

Depending on application the recognition problem may be more or less sophisticated. If automatic speech recognition is applied to diagnosis case the number of possible substitution and deformation of the particular phoneme is large and recognition task is relatively difficult. Moreover recognized substitutions and deformations are usually acoustically and phonetically similar what additionally complicates the problem. Another problem is speaker independency in diagnosis recognition task.

Automatic speech recognition used during therapy is much simpler problem. In this case the kind of speech disorder is known and usually recognition task is limited to recognition of the two phonemes from closed set. First phoneme is correct realization and usually is recognized in speaker independent manner. The second phoneme is bad realization and can be recognized in speaker dependent manner. In this paper we mostly investigate the substitutions and assume that recognition is applied to the therapy case.

Since selection of discriminating signal features is a crucial issue in speech recognition tasks, in the present paper, application of recently proposed Human Factor Cepstral Coefficients [4] to automatic recognition of pathological phoneme pronunciation in speech of impaired children is tested and compared in efficiency to usage of standard mel-frequency cepstral coefficients (MFCC) as a feature vector. Both dynamic time warping (DTW), working on whole words or embedded phoneme patterns, and hidden Markov models (HMM) are used as classifiers in the presented research. Obtained results demonstrate superiority of combining HFCC features and modified phoneme-based DTW classifier.

2. Methods

2.1. Feature extraction methods

2.1.1. Mel-frequency cepstral coefficients (MFCC)

The *mel-frequency cepstral coefficients* (MFCC), originated from modeling of acoustic signal processing performed in cochlea. Detailed description of the features can be found in [3], [5]. The important characteristics of MFCC is equally spacing of centers of triangular frequency filters along mel-frequency scale.

2.1.2. Human-factor cepstral coefficients (HFCC)

The novel *human factor cepstral coefficients* (HFCC) approach to speech features extraction has been proposed and described in details in [4]. The method and its algorithmic implementation are very similar to the MFCC. The only but crucial difference between these two methods is that now filter bandwidth is decoupled from filter spacing. In HFCC filter center frequencies are equally spaced in mel frequency scale (1), as in the MFCC method, but filter bandwidth is a design parameter, measured in equivalent rectangular bandwidth (ERB):

$$ERB = 6.23 f_c^2 + 93.39 f_c + 28.52 \quad \text{Hz} \qquad (1)$$

where filter center frequency f_c is expressed in kHz. When wider filter bandwidth than ERB is exploited (ERB scaled by some factor > 1) then the HFCC-based speech recognition can be under some circumstances more resistant to noise. Further details concerning the method can be found in [4], [6].

2.2. Recognition methods

Recognition or classification methods should be properly selected to the given speech recognition task. Here, the method should recognize phoneme embedded in the word with assumption that word comes from the closed set of two classes. The first class is correct word realization and the second one is incorrect. Both word classes can be distinguished only by one phoneme. In order to accomplish the recognition, four methods were considered: word-based Dynamic Time Warping (DTW), phoneme-based DTW and Hidden Markov Models (HMM) with whole word models and HMM with phoneme models.

2.2.1. Word-based dynamic time warping

Dynamic time warping (DTW), as good speech classifier, is well known for many years. It was used in reported research for its simplicity of implementation and analysis as well as for relatively high recognition accuracy comparable with HMM method.

An Euclidean distance was used as a distance measure in the reported research. An accumulated distance at each point of the search path was calculated according to recursive procedure given by the equations:

$$g(i,j) = \min \begin{bmatrix} g(i-2,j-1) + d(i,j) \\ g(i-1,j-1) + d(i,j) \\ g(i-1,j-2) + d(i,j) \end{bmatrix} \quad (2)$$

In order to normalize obtained result the accumulated cost was divided by factor D:

$$D = \sqrt{N_w^2 + N_S^2} \quad (3)$$

where: N_W – number of feature vectors of the reference pattern, N_S – number of feature vectors of the word being recognized. The search path was limited by two parallel lines shifted by the coefficient:

$$Q = round(w \cdot \max(N_S, N_W)) \quad (4)$$

where: w – path width coefficient (equals 0.2 in the reported research). A detailed description of DTW algorithm can be found in [2].

2.2.2. Phoneme-based recognition by DTW method

Standard DTW method described in the previous chapter is based on whole word models. Such approach is suitable for recognizing isolated words especially significantly differing from each other. When the words differ only by one short speech segment, for instance one phoneme, word-based approach often fails particularly when the phonemes distinguishing two words are acoustically similar. Moreover, segments outside distinguishing phonemes are addicted to disturbances like variations in speaking style, another mispronounced phoneme or external noise what can give higher global DTW distance beetwen the words of the same class.

Proposed solution assumes that class of the recognized word is known and that this word can be spoken correctly or incorrectly in the earlier diagnosed manner. Two realisation of the word can be distinguished only on one phoneme position. For example Polish word *szafa* [ʃa:fa] can be incorrectly pronounced like [*sa:fa*] so these two words can be distinguished by the first phoneme only.

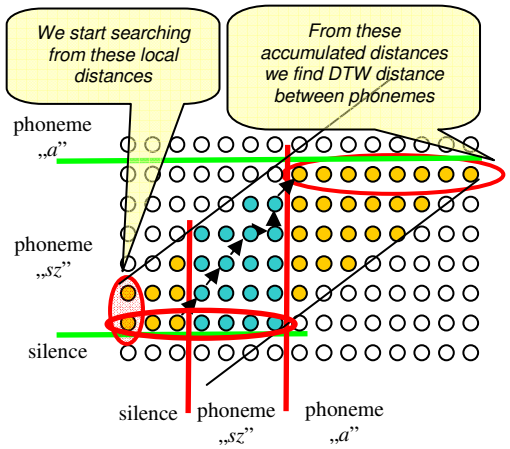

Figure 1: Application of DTW to the recognition of utterance segment

Proposed method can be implemented in the following way:

Before recognition the words coming from training set should be segmented. It is sufficient to match the segment being potentially a mispronounced phoneme.

The start and the end region of the recognized phoneme is determined with local distance array (Figure 1).

The standard DTW procedure starts from the start region and ends in the end region (Figure 1).

Research showed that proper segmentation of the speech patterns is crucial for high accuracy phoneme recognition. This segmentation should be done not only by ear or word waveform observation but also by spectrogram analysis.

2.2.3. Hidden Markov Models based recognition

The third classification methods investigated in the research was Hidden Markov Models. Both HMMs with the phoneme and whole word as a modeled unit were used. Before classifying by the HMM, training procedure has been performed.

In whole word training models re-estimation by using Viterbi algorithm and Baum-Welch algorithm were used. In the recognition process Viterbi algorithm was used.

Initialization of the HMMs Parameters: A_T, μ_j, Σ_j
Embedded Trainnig x 3
Fixing the Silence Models 3
Embedded Trainnig x 2
Making Triphones from Monophones
Embedded Trainnig x 2
Making Tied-State Triphones
Embedded Trainnig x 2
Finally estimated HMM parameters: A_T, μ_j, Σ_j

Figure 2: *Block diagram of the training procedure in case of phoneme based HMMs.*

In case of training of phoneme HMMs, a fundamental training procedure was so called embedded training modified Baum-Welch algorithm. Detailed description of the embedded training is beyond scope of this paper. Information on this procedure can be found in [7]. Complete training process diagram is shown in Figure 2. As a hidden Markov model for phoneme simple three-state left-right model with no skips was used.

For recognition procedure in phoneme HMMs an alternative formulation of the Viterbi algorithm was used called the *Token Passing Model*. Comprehensive description on the token passing algorithm can be found in [8].

3. Experiments and Results

3.1. Finding an optimal classifier (recognizer)

3.1.1. Research methodology

In the experiments the recognized speech utterances were the following pairs of Polish phonemes:

- {*s*, *sz*}, {si, *sz*} extracted from word ***sz**afa* [**ʂ**a:fa] and its deformed versions: ***s**afa* [**s**a:fa], ***si**afa* [**ɕ**a:fa]
- {*c*, *cz*}, {ci, *cz*} extracted from word ***cz**apka* [**t͡ʂ**a:pka] and its deformed versions: ***c**apka* [**ts**a:pka], ***ci**apka* [**t͡ɕ**a:pka]
- {*dz*, *drz*}, {*dzi*, *drz*} extracted from word ***drz**ewo* [**d͡ʐ**e:vo] and its deformed versions: ***dz**ewo* [**dz**e:vo], ***dzi**ewo* [**d͡ʑ**e:vo]

Among the recorded words spoken by impaired children there were records with substituted phonemes. However their number was insufficient for research purpose. Therefore the set of recordings was supplemented by the records coming from persons who imitated wrong pronunciation of tested words. There were 6÷8 samples per word in a training set and 12÷16 samples per word in a testing set. Detailed structure of the sets was presented in our former research [9]. The speech samples were recorded with 48 kHz sampling rate and 16 bits/sample. Relatively high sampling frequency has been chosen in order to not attenuate spectral components which can be significant with regard to recognition process.

3.1.2. Phoneme-based vs. word-based DTW classifier

Recognition experiments were carried out with various combinations of the MFCC and HFCC parameters. Basic parameters are presented in Table 1

Table 1. *Basic parameters of the feature extraction procedures*

Parameter	Features	
	MFCC	HFCC
Preemphasis $H(z)$	$1–0.9375z^{-1}$	$1–0.9375z^{-1}$
Frame length	30 ms	30 ms
Frame shift	10 ms	10 ms
Window	Hamming	Hamming
No of Filters	30	30
Number of coefficients	15	15
DFT Length	4096	4096
ERBscaleFactor	--	1 ÷ 6

In the research phoneme based and standard (word based) DTW procedure was used. The best achieved results for DTW method are presented on Figure 3. In Table 2 mean recognition accuracies for phoneme based (PB) and word based (WB) method was presented. Besides basic parmeters (PP) Delta coefficients (D) and Delta-delta coefficients (DD) were added during subsequent experiments.
It can be observed from Table 2 that:

- Average recognition accuracy is higher for MFCC features in comparison with HFCC features providing word-based DTW method is used.
- Average recognition accuracy is higher for HFCC features in comparison with MFCC features providing phoneme-based DTW method is used.
- Statistically phoneme-based DTW method gives better recognition results than word-based DTW method.

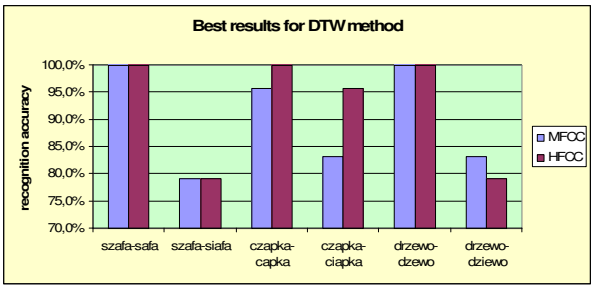

Figure 3: *Comparison of MFCC based and HFCC based DTW method*

Table 2. *Mean recognition accuracies for phoneme based (PB) and word based (WB) DTW method.*

	MFCC		HFCC_1		HFCC_2		Average	
	WB [%]	PB [%]	WB [%]	PB [%]	WB [%]	PB [%]	WB [%]	PB [%]
PP	84,7	85,4	81,9	86,8	84,0	88,9	83,5	87,0
D	83,3	86,8	82,6	88,2	80,6	87,5	82,2	87,5
DD	84,7	81,2	81,3	84,0	81,3	87,5	82,4	84,2
Avrg.	84,2	84,5	81,9	86,3	82,0	88,0	82,7	86,3

3.1.3. HMM clasifiers

The first one HMM classifier examined in the research was classifier based on whole word HMMs. Dependency of the word recognition accuracy on number of HMMs was observed. The rules which should be used to choose the optimal state number are presently unclear. It could be only stated that there is weak negative correlation beetwen state number and number of the utterance phonemes for basic MFCC parameters and there is lack of correlation for delta and delta-delta coefficients as additional feature sets. It could be stated also that most frequent values indicating optimal state number per one phoneme ranged from 2 to 3 emitting states per one phoneme.

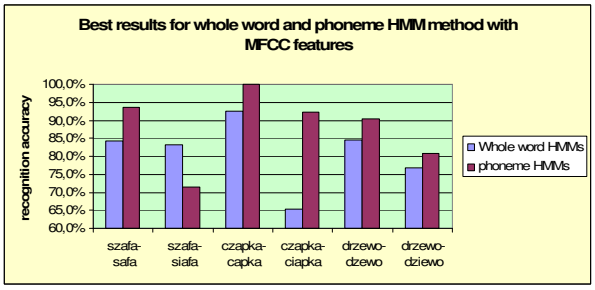

Figure 4: *Comparison of recognition accuracies achieved by whole word HMMs and phoneme HMMs.*

The phoneme HMMs based on MFCC were investigated after examining whole word HMMs. Comparison of the two methods is presented in Figure 4. Much better results have been achieved using phnoneme HMMs. Therefore this method has been chosen for further research.

Besides of MFCC, the HFCC features were also used in recognition experiments with HMMs of the phoneme. The best achieved results are presented in Figure 5.

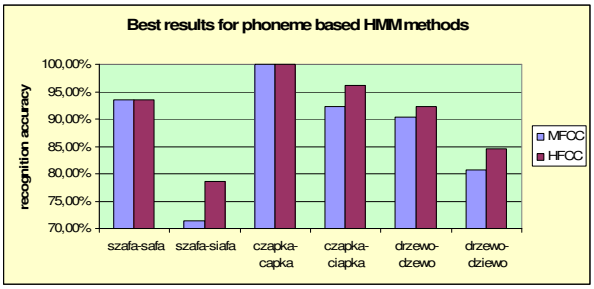

Figure 5: *Comparison of recognition accuracies for phoneme HMMs method with MFCC and HFCC features used.*

Mean recognition accuracies for phoneme based HMMs with HFCC features versus ERB scale factor were depicted on Figure 6. It can be observed that the best results have been achieved for ERB factor values ≥ 1. It is also evident that best results are for delta-delta coefficients added.

As a reference points mean MFCC recognition accuracy was also calculated. The results were as follows:

- 80.27 % for basic parameters
- 82.63 % for delta coefficients added
- 86.93 % for delta-delta coefficients added

These results are from 2.16 % up to 5.13 % worse in comparison with the best mean accuracies obtained by HFCC method.

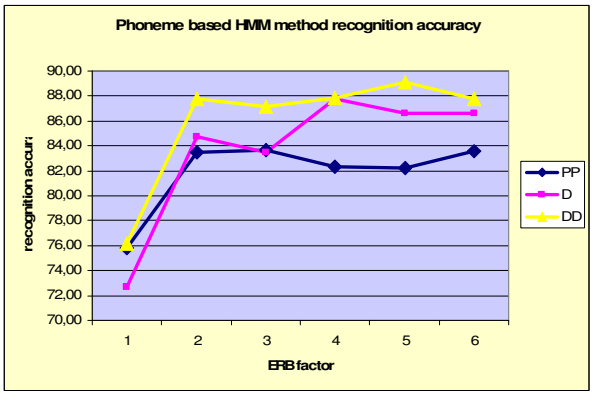

Figure 6: *Mean recognition accuracies for phoneme HMMs with different parameters of HFCC features used.*

Figure 7 presents comparison results for DTW and HMM methods for HFCC features. Decision on using proper classifier depends on recognition task for example in case *drzewo-dzewo* DTW gives better results but for *drzewo-dziewo* better results were obtained with HMM. Mean recognition accuracy was 1,5 % higher for DTW method.

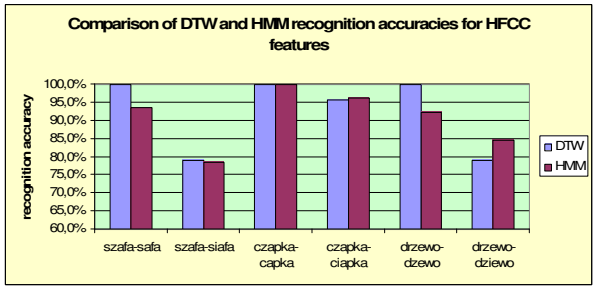

Figure 7: *Comparison of DTW and HMM methods.*

4. Conclusions

In the paper comparative analysis of using DTW and HMM methods in pathological speech recognition has been presented. The therapy case was considered. Obtained results indicate that HFCC based speech recognition gives better results in comparison with standard MFCC features and after proper selection of parameters can be used for pathological speech recognition task. At the present stage of research it can be stated that DTW method outperforms HMM ones in considered recognition task, however further problem investigation is necessary. More phonemes will be tested and larger sets of speech samples will be used in future research.

Obtained results allow to implement recognition methods in real word application for the therapy of substitution: *sz-s*, *cz-c*, *drz-dz*. For substitution *sz-si*, *cz-ci*, *drz-dzi* more efficient methods have to be found. Probable research direction could be PCA and discriminant analysis.

5. Acknoledgement

The work was sponsored by Polish Scientific Committee from KBN grant no. 1 H01F 046 28.

6. References

[1] J. T. Kania, Foundations of the linguistic classification of speech disorder (Podstawy językoznawczej klasyfikacji zaburzeń mowy), [in:] Szkice logopedyczne, Lublin, PTL 2001, pp. 11-30 (in Polish).

[2] Rabiner LR, Juang BH: Fundamentals of Speech Recognition. Prentice Hall 1993.

[3] Picone JW: Signal modeling techniques in speech recognition. Proc IEEE, 81(9): 1215-1247, 1993.

[4] Skowronski MD, Harris JG: Exploiting independent filter band-width of human factor cepstral coefficients in automatic speech recognition. J. Acoust. Soc. Am., 116(3): 1774-1780, 2004.

[5] VoiceBox: Speech Processing Toolbox for Matlab, http: //www.ee.ic.ac.uk/hp/staff/dmb/voicebox/ voicebox.html.

[6] HFCC, http://www.cnel.ufl.edu/~markskow/.

[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book (for HTK Version 3.0), http://htk.eng.cam.ac.uk, Jul. 2000.

[8] S. J. Young, N. H. Russell , J. H. S. Thornton, Token Passing: a Conceptual Model for Connected Speech Recognition Systems, CUED Technical Report F_INFENG/ TR38, Cambridge University, 1989.

[9] R. Wielgat, T. Zieliński, Ł. Hołda, D. Król, T. Woźniak, S. Grabias, HFCC Based Pathological Speech Recognition, AQL, Gronningen, Netherlands, Oct. 2006.