Automatic Recognition of Pathological Phoneme Production

Robert Wielgat^a Tomasz P. Zieliński^b Tomasz Woźniak^c Stanisław Grabias^c
Daniel Król^a

^aDepartment of Technology, Higher State Vocational School in Tarnów, Tarnów, ^bDepartment of Telecommunications, AGH University of Science and Technology, Krakow, and ^cDivision of Logopedics and Applied Linguistics, Maria Curie-Skłodowska University, Lublin, Poland

Key Words

Therapy of speech disorders • Automatic speech recognition • Dynamic time warping • Hidden Markov models • Human-factor cepstral coefficients

Abstract

Objective: Proper diagnosis and therapy of pathological pronunciation of phonemes play an important role in modern logopedics. To enhance the efficiency of diagnosis and therapy an automatic recognition of pathological phoneme pronunciation is addressed in this paper. The authors focus on the therapy of phoneme substitution disorders. **Patients and Methods:** Recognized speech samples come from speech-impaired Polish children and partially from persons imitating speech disorders. Recognized speech disorders were substitutions in pairs {s, ʃ}, {ɛ, ɕ}, {ts, tɕ}, {tɕ, tɕ}, {dɕ, dɕ}, and {dɕ, dɕ} embedded in Polish carrier words. In order to detect substitutions in the recognized words, recently proposed human factor cepstral coefficients (HFCC) have been implemented. Efficiency of the HFCC approach was compared to the application 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) were used as classifiers. The HMM classifier was based on whole-word models as well as phoneme models. Results present a comparative analysis of DTW and HMM methods. **Conclu-**

sions: The superiority of HFCC features over those of MFCC was demonstrated. Results obtained by DTW methods, mainly by modified phoneme-based DTW classifier, were slightly better in comparison with the HMM classifier. Results obtained for the detection of substitution in pairs {s, ʃ}, {ts, tɕ}, {dɕ, dɕ} are very promising. The methods developed for these cases can be integrated into computer systems for speech therapy. For substitutions in pairs {ɛ, ɕ}, {tɕ, tɕ}, {dɕ, dɕ} further research is necessary.

Copyright © 2008 S. Karger AG, Basel

Introduction

Automatic speech recognition can significantly improve the diagnosis and therapy of the most common speech disorders. The picture questionnaire is an elementary type of logopedic assessment. A person (child) is asked to say the names of pictures visible on cards or appearing on the computer screen. The answers given form the basis for identifying wrong pronunciations of words and for defining the type of speech disorder. When a mispronounced word is also recognized by the computer, accuracy and efficiency of the diagnosis can increase significantly.

Pathological pronunciation of sounds typically occurs in paradigmatic speech disorders. Using a linguistic classification, paradigmatic disorders are divided into: (1) eli-

Fig. 1. Automatic speech recognition task in logopedic diagnosis and therapy for the Polish word *szafa* [ʂa:fa].

sion (no realization of phoneme), (2) substitution (realization of phoneme replaced by realization of other phonemes), and (3) deformation. This classification is applicable to all European languages. The speech recognition task in this case concerns the recognition of selected phonemes embedded in the utterance.

Recognizing phonemes automatically in the case of logopedical diagnosis is rather difficult because the number of possible substitutions and deformations of a particular phoneme is relatively large. Moreover recognized substitutions and deformations are usually acoustically and phonetically similar, which additionally complicates the problem. Another difficulty is speaker independence.

Automatic speech recognition used during therapy is much simpler. In this case, the kind of speech disorder is known and usually the recognition task is limited to recognition of two phonemes from a closed set. The first phoneme is the correct realization and is usually recognized in a speaker-independent manner. The second phoneme is a bad realization and can be recognized in a speaker-dependent manner. A schematic explanation of recognition tasks in the diagnosis and therapy of the Polish word *szafa* [ʂa:fa] is presented in figure 1.

In the present paper we mostly investigate sound substitutions and assume that recognition is applied to the therapy case. Effective speech recognition methods capable of detecting substitutions can be integrated into

various computer games, making speech therapy much more attractive for children. We focus our attention on the most commonly disturbed Polish sounds, i.e. realizations of dentalized phonemes, especially ʂ, ʐ, tʂ, dʂ (80% of a population of Polish children with pronunciation disorders produce these phonemes incorrectly).

From the automatic speech recognition point of view any recognition task consists basically of two stages: feature extraction and classification. Feature extraction is a procedure which processes a signal into a sequence of feature vectors. Selection of discriminating signal features is a crucial issue at this stage of processing. In the present paper, application of the recently proposed human factor cepstral coefficients (HFCC) [2] to automatic recognition of pathological phoneme pronunciation in children's speech is tested and compared in efficiency to standard mel-frequency cepstral coefficients (MFCC) as a feature vector.

The recognition or classification methods selected should be adequate for the given speech recognition task. Here, the method should recognize a phoneme embedded in an word on assumption that the word comes from the closed set of two classes. The first class is the 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, three methods were considered: word-based dynamic time warping

Table 1. Basic parameters of the feature extraction procedures

Parameter	Features	
	MFCC	HFCC
Preemphasis $H(z)$	$1 - 0.9375 z^{-1}$	$1 - 0.9375 z^{-1}$
Frame length, ms	30	30
Frame shift, ms	10	10
Window	Hamming	Hamming
Filters, n	30	30
Coefficients, n	15	15
DFT length	4,096	4,096
ERB scale factor	–	1–6

DFT = Discrete Fourier Transform.

(DTW), phoneme-based DTW and hidden Markov models (HMM).

The DTW method calculates a distance between two feature vector sequences: pattern sequence and recognized sequence. This distance is called a DTW distance. A class of patterns with minimal DTW distance is assigned to the recognized word. Two DTW methods were tested in this study. The first one was the standard DTW method where the pattern sequence of feature vectors comes from the whole word. In the second method we introduced a modification: instead of using DTW distance for whole-word patterns a DTW distance between embedded phonemes of pattern and recognized word was calculated.

The second type of classifier was HMM. In the HMM method a probabilistic model for particular speech units is being built during the training procedure. The modeled speech unit is usually the whole word or the phoneme. The speech units are also combined to create a so-called word network. At the recognition stage the most probable model is chosen. A class associated with the most probable model is a recognized class. The HMM for whole words and for phonemes were tested in this study. Two different concepts of word networks were also examined.

Methods

Feature Extraction via MFCC

Calculation of the MFCC, originating from a model of acoustic signal processing performed in the cochlea, was performed in the following steps [3, 4]:

- 1 blocking signal into frames and windowing by Hamming window;

- 2 calculation of the fast Fourier transform (FFT) in the windowed frames;
- 3 packing the FFT power into the uniform, overlapping by 50% mel frequency bands with equally spaced center mel frequencies using triangular weighting in mel-scale (number of bands is a parameter of the algorithm); conversion from linear-frequency scale to the mel-frequency scale and vice versa is given by the equations:

$$f_{mel} = 2,595 \log_{10}(1 + f_{Hz}/700), f_{Hz} = 700 \cdot (10^{f_{mel}/2,595} - 1)$$

- 4 calculation of log spectral power coefficients in mel bands;
- 5 performing DCT on coefficient vectors ($n = 0, 1, \dots, q - 1$):

$$X(n) = c(n) \sum_{k=0}^{K-1} \ln(S_k) \cos\left(\frac{\pi(2k+1)n}{2K}\right),$$

$$c(0) = \sqrt{\frac{1}{K}}, \dots, c(n) = \sqrt{\frac{2}{K}}$$

where S_k = log spectral power coefficient in the k -th frequency band; K = number of frequency bands; q = number of MFCC coefficients;

- 6 calculating first and second derivatives of the MFCC coefficients with respect to time, so-called delta and delta-delta coefficients, respectively.

Feature Extraction via HFCC

The novel HFCC approach to speech feature extraction has been proposed and described in detail by Skowronski and Harris [2]. The method and its algorithmic implementation are very similar to the MFCC described above. 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 the mel-frequency scale [1], as in the MFCC method, but filter bandwidth is a design parameter, measured in equivalent rectangular bandwidth (ERB):

$$ERB = 6.23f_c^2 + 93.39f_c + 28.52 \text{ Hz}$$

where filter center frequency f_c is expressed in kilohertz. When a wider filter bandwidth than ERB is exploited (ERB scaled by some factor >1) then HFCC-based speech recognition can be more resistant to noise under some circumstances. Moreover such a signal representation is much closer to signal processing in the cochlea than MFCC. Further details concerning the method can be found in Skowronski and Harris [2, 5].

Feature Extraction Parameters and Values

Recognition experiments were carried out with various combinations of MFCC and HFCC parameters. Their basic values are presented in table 1. Besides basic parameters also delta and delta-delta coefficients were optionally used during subsequent experiments (first and second derivative of MFCC/HFCC coefficients with respect to time).

Classification via Word-Based DTW

DTW has been known to be a good speech classifier for many years. In the present study it was used for its simplicity of implementation and analysis as well as for its relatively high recognition accuracy, comparable with the HMM method.

Fig. 2. Optimal search path restricted by two parallel lines in DTW algorithm.

The main idea of using the DTW algorithm for speech recognition is finding an optimal path with minimal cost from the lower left corner to the upper right corner of the local distance array (fig. 2). A single element d_{mn} of the array equals the distance between the m -th feature vector (for instance MFCC or HFCC) of a recognized utterance and the n -th feature vector (MFCC or HFCC) of the reference pattern. A 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 a special recursive procedure originating from dynamic programming methods. Normalized accumulated distance at the end of a search path is a DTW distance representing the measure of similarity between two words represented as sequences of feature vectors. The smaller the DTW distance, the more similar the compared words. In order to limit a search region various limitations are introduced, for example two parallel lines visible in figure 2. Detailed description of the DTW algorithm can be found in Rabiner and Juang [6], Sakoe and Chiba [7], and Kuhn and Tomaszewski [8].

Classification via Phoneme-Based DTW

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

The proposed solution assumes that the class of the recognized word is known and that this word can be spoken correctly or incorrectly in the earlier diagnosed manner. Two realizations of the word can be distinguished only at one phoneme position. For example, the Polish word *szafa* [ʂa:fa] can be incorrectly pronounced as [sa:fa], so that these two words can be distinguished by the first phoneme only. The method can be implemented in the following way:

- 1 Before recognition, the words from the training set should be segmented. It is sufficient to match the segment that is a potentially mispronounced phoneme.
- 2 The start and the end region of the recognized phoneme is determined with a local distance array (fig. 3).
- 3 The standard DTW procedure starts from the start region and ends at the end region (fig. 3).

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

Classification via HMM

The third classification method investigated was HMM. Both HMMs with the phoneme and the whole word as a modeled unit were used. Before classifying by the HMM, a training procedure was performed.

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

In case of training of phoneme HMMs, the fundamental training procedure was the so-called embedded training modified Baum-Welch algorithm. Detailed description of the embedded training is beyond the scope of this paper. Information on this procedure can be found in Young et al. [9]. A diagram of the complete training process is shown in figure 4. As an HMM for phonemes a simple three-state left-right model with no skips was used. The probability of observations both for HMMs of phonemes as well as for HMMs of whole words was single Gaussian. Two word networks were tested: a whole-word network and a phoneme word network (fig. 5).

For the recognition procedure in phoneme HMMs an alternative formulation of the Viterbi algorithm was used, called the Token Passing Model. A comprehensive description of the token passing algorithm can be found in Young et al. [10].

Research Methodology

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

- 1 {s, ʂ}, {ɛ, ʂ} extracted from the word *szafa* [ʂa:fa] and its deformed versions: *safa* [sa:fa], *siafa* [ɛa:fa],
- 2 {ts, tʂ}, {tɛ, tʂ} extracted from the word *czapka* [tʂa:pka] and its deformed versions: *capka* [tsa:pka], *ciapka* [tɛa:pka], and
- 3 {dz, dʒ}, {dɛ, dʒ} extracted from the word *drzewo* [dʒe:vo] and its deformed versions: *dzewo* [dze:vo], *dziewo* [dɛe:vo].

Among the recorded words produced by children with speech impairment there were records with substituted phonemes. However, their number was insufficient for research purposes. Therefore the set of recordings was supplemented by records coming from persons who imitated wrong pronunciations of the test words. There were 6–8 samples per word in a training set and 12–16 samples per word in a testing set.

The detailed structure of the sets is presented in table 2. The speech samples were recorded at 48 kHz sampling rate and 16 bits/sample with a standard capacitive microphone. A relatively high sampling frequency was chosen in order to not attenuate spectral components that can be significant with regard to the recognition process.

Fig. 3. Application of DTW to the recognition of an utterance segment.

Table 2. Structure of the training and testing sets used in the experiments

Experiment No.	Word pronunciation	Type of set			
		training set		testing set	
		female	male	female	male
1	szafa	4	4	8	8
	safa	4	4	7	9
2	szafa	3	3	6	6
	siafa	3	3	6	6
3	czapka	3	4	6	7
	capka	3	4	7	6
4	czapka	3	4	6	7
	ciapka	3	4	6	7
5	drzewo	3	4	6	7
	dzewo	3	4	6	7
6	drzewo	3	4	6	7
	dziewo	3	4	6	7

Results

Phoneme-Based vs. Word-Based DTW Classifiers

In the first experiment designed to investigate the recognition of substitutions, phoneme-based and standard (word-based) DTW procedures were used. The best results achieved with the DTW method are presented in figure 6. Mean recognition accuracies for phoneme-based and word-based methods are presented in table 3. These

Fig. 4. Block diagram of the training procedure in case of phoneme-based HMMs.

mean accuracies were calculated as the average of the best accuracies in each of the six experiments from table 3. Recognition was performed for basic MFCC and HFCC parameters as well as for delta coefficients and delta-delta coefficients. It can be observed from table 3 that:

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

Fig. 5. Word networks used in experiments. **a** Whole-word network. **b** Phoneme word network.

Fig. 6. Comparison of MFCC-based and HFCC-based DTW methods.

Fig. 8. Comparison of recognition accuracies for phoneme HMM method with MFCC and HFCC features used.

Fig. 7. Comparison of recognition accuracies achieved by whole-word HMMs and phoneme HMMs.

HMM Classifiers

The first HMM classifier examined in the present study was a classifier based on whole-word HMMs. Dependency of the word recognition accuracy on the number of HMMs was observed. The rules that should be

Table 3. Mean recognition accuracies (%) for phoneme-based (PB) and word-based (WB) DTW methods

	MFCC	HFCC
WB	86.81	86.81
FB	89.58	91.67

used to choose the optimal state number are presently unclear. It can only be stated that there is a weak negative correlation between the number of states and the number of phonemes in an utterance for basic MFCC parameters and that there is a lack of correlation for delta and delta-delta coefficients as additional feature sets. It could also be stated that the most frequent values indicating optimal state number per phoneme ranged from 2 to 3 emitting states per phoneme. The phoneme HMMs based on MFCC were investigated after examining whole-word HMMs. Comparison of the two methods is presented in figure 7. Much better results were achieved using pho-

neme HMMs. Therefore this method has been chosen for further research. Besides MFCC, HFCC features were also used in recognition experiments with HMMs of phonemes. The best achieved results are presented in figure 8.

Mean recognition accuracies for phoneme-based HMMs with HFCC features versus ERB scale factor are depicted in figure 9. It can be observed that the best results were achieved for ERB scale factor values ≥ 1 . It should be mentioned that there are different optimal ERB scale factors for different pairs of phonemes. More detailed analysis of the problem was presented in our previous research [11, 12]. It is also evident from figure 9 that best results are obtained for delta-delta coefficients added.

As a reference point, mean MFCC recognition accuracy was also calculated. The results were as follows: 80.27% for basic parameters, 82.63% for delta coefficients added, and 86.93% for delta-delta coefficients added. These results are from 2.16 to 5.13% worse in comparison with the best mean accuracies obtained by HFCC method.

As for word networks better results were achieved using a whole-word network in contrast to a phoneme word network (fig. 10). Figure 11 presents a comparison of results for DTW and HMM methods for HFCC features. Decision on the proper classifier depends on the recognition task; for example, for *drzewo-dzewo* DTW gives better results, but for *drzewo-dziewo* better results were obtained with HMM. Mean recognition accuracy was 1.5% higher for the DTW method.

Discussion

Although the results obtained in the present study are ambiguous to some extent and are not statistically significant, they indicate some trends among the speech data and are very promising as for real-world applications.

The first trend that can be observed is an overall superiority of phoneme-based methods over whole-word-based ones in the detection of substituted phonemes in different words. This means that the phoneme-based DTW method and the phoneme-based HMM method gave the best recognition results in comparison with whole-word DTW and whole-word HMMs. Explanation of this phenomenon in the case of the DTW method is rather simple. DTW distance in case of the whole-word-based method is strongly influenced by phonemes being not significant in the recognition process. For example,

Fig. 9. Mean recognition accuracies for phoneme HMMs with different parameters of HFCC features used. PP = Basic parameters; D, DD = delta and delta-delta coefficients added.

Fig. 10. Comparison of HMM recognition accuracies for HFCC features with different word networks.

Fig. 11. Comparison of DTW and HMM methods.

Fig. 12. Spectrograms of Polish phonemes. **a** Phoneme ɕ. **b** Phoneme ʑ. **c** Phoneme s.

in the word *szafa* it is necessary to recognize only the first phoneme, the rest of phonemes is of no use in recognition and may cause unnecessary increase in DTW distance. So the phoneme-based DTW recognition, where only the phoneme under consideration is recognized, usually outperforms the whole-word-based DTW method where DTW distance is created by aligning all the phonemes in the word.

Overall superiority of phoneme-based HMMs over whole-word HMMs is not clear at the present stage of work. A probable explanation of this trend is that HMMs for some phonemes which are repeated in the word, for instance phoneme *a* in our case, can be better trained than phonemes that are not repeated. It is due to parameter tying in the HMM training process. Therefore calculated probability of the model during recognition is greater. In case of whole-word models parameter tying is not performed, so particular HMM parameters are not really well trained. It is also not clear why a whole-word network is better than a phoneme one.

Another trend observed in the experiments is HFCC superiority over the MFCC. This finding is consistent with our former results [11, 12], and with research obtained for English speech recognition [2] as well as with the results we obtained for the recognition of bird voices [13]. In the case of substitution detection, recognition accuracy can be additionally enhanced by choosing optimal ERB scale factors for different phoneme pairs. HFCC superiority over MFCC is explained by a closer relationship between HFCC features and hearing physiology in comparison with MFCC [2].

As shown in figure 11, the best recognition accuracies were achieved for substitutions {s, ɕ}, {ts, tɕ}, and {dz, dɕ}

(words *szafa-safa*, *czapka-capka*, *drzewo-dzewo*) (100%), while for substitutions {ɕ, ɕ}, {tɕ, tɕ}, and {dɕ, dɕ} (words *szafa-siafa*, *czapka-ciapka*, *drzewo-dziewo*) they were up to 21% worse. This phenomenon can be explained by analyzing tongue position during pronunciation. The end of the tongue touches the palate in case of correct phonemes ɕ, tɕ, dɕ; it touches the base of the upper teeth in case of phonemes ʑ, tʑ, dʑ, and the lower teeth in case of phonemes s, ts, dz. Thus differences in tongue position are smaller in phonemes {ɕ, ɕ}, {tɕ, tɕ}, {dɕ, dɕ} than in phonemes {s, ɕ}, {ts, tɕ}, {dz, dɕ}. Smaller differences in tongue position make the phonemes more similar to each other. This similarity can be observed in real-world observations. For example, for a native speaker of English it is hard to distinguish the Polish phonemes ɕ and ɕ (both pronunciation and listening), but it is easier to distinguish the phonemes s and ɕ where the difference in tongue position is bigger and phonemes are less similar to each other. Phonetic similarity of phonemes is also evident in the frequency spectrum (fig. 12). For instance, the main spectral energy of phonemes ɕ and ɕ lays in the range 2–10 kHz while for the phoneme s this range is 4–13 kHz. Hence better recognition results for substitution {s, ɕ} in comparison with substitution {ɕ, ɕ}. Similar dependences can be observed for the rest of phonemes examined in this study.

In conclusion, the methods explored in this study hold great promise for the application to computer programs (for instance computer games) for the therapy of selected substitutions. For the phonemes recognized with lower accuracy it is necessary to develop more efficient methods. The main direction of future research can be further optimizing HMM parameters, using dis-

criminant analysis or PCA. In order to obtain more reliable and statistically relevant results, experiments with a larger number of phoneme types and a larger number of samples should be carried out. It would also be very interesting to apply the method for the recognition of speech disorders to languages different from the Polish language.

Acknowledgments

The authors wish to thank the Polish Ministry of Science and Higher Education for supporting this study. The study was supported by the grant of the Polish Ministry of Science and Higher Education no. 1 H01F 046 28.

References

- 1 Kania JT: Foundations of the linguistic classification of speech disorders (in Polish); in *Szkice logopedyczne*. Lublin, PTL, 2001, pp 11–30.
- 2 Skowronski MD, Harris JG: Exploiting independent filter band-width of human factor cepstral coefficients in automatic speech recognition. *J Acoust Soc Am* 2004;116:1774–1780.
- 3 VoiceBox: Speech Processing Toolbox for Matlab, <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.
- 4 Picone JW: Signal modeling techniques in speech recognition. *Proc IEEE* 1993;81: 1215–1247.
- 5 HFCC, <http://www.cnel.ufl.edu/~markskow/>.
- 6 Rabiner LR, Juang BH: *Fundamentals of Speech Recognition*. Englewood Cliffs, Prentice-Hall, 1993.
- 7 Sakoe H, Chiba S: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Acoust, Speech Signal Processing* 1978;ASSP-26:43–49.
- 8 Kuhn MH, Tomaschewski HH: Improvements in isolated word recognition. *IEEE Trans Acoust Speech Signal Processing* 1983;ASSP-31:157–167.
- 9 Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P: *The HTK Book* (for HTK version 3.0). <http://htk.eng.cam.ac.uk>, July 2000.
- 10 Young SJ, Russell NH, Thornton JHS: Token passing: a conceptual model for connected speech recognition systems. *CUED Tech Rep F_INFENG/TR38*. Cambridge University, 1989.
- 11 Wielgat R, Zieliński TP, Hołda Ł, Król D, Woźniak T, Grabias S: HFCC based pathological speech recognition. 7th Int Conf Advances in Quantitative Laryngology, Voice and Speech Research, AQL-2006, Groningen, 2006.
- 12 Zieliński TP, Gajda P, Stachura M, Wielgat R, Król D, Woźniak T, Grabias S: Zastosowanie współczynników HFCC jako cech sygnału mowy w automatycznej detekcji wad wymowy. *MiSSP'2006, XVI Sympozjum Modelowanie i Symulacja Systemów Pomiarowych*, Krynica, września 2006.
- 13 Wielgat R, Zieliński TP, Potempa T, Lisowska-Lis A, Król D: HFCC-based recognition of bird species. *Signal Processing 2007, Scientific Workshop*, Poznań, September 2007.