HMM, recognition,
speech, disorders

Marek WIŚNIEWSKI*, Wiesława KUNISZYK-JÓŹKOWIAK*,
Elżbieta SMOŁKA*, Waldemar SUSZYŃSKI*

AUTOMATIC DETECTION OF PROLONGED FRICATIVE PHONEMES WITH THE HIDDEN MARKOV MODELS APPROACH

The Hidden Markov Model (HMM) is a stochastic approach to recognition of patterns appearing in an input signal. In the work author's implementation of the HMM were used to recognize speech disorders - prolonged fricative phonemes. To achieve the best recognition effectiveness and simultaneously preserve reasonable time required for calculations two problems need to be addressed: the choice of the HMM and the proper preparation of an input data. Tests results for recognition of the considered type of speech disorders are presented for HMM models with different number of states and for different sizes of codebooks.

1. INTRODUCTION

The HMMs are stochastic models and are widely used for recognition of various patterns. They gained great significance particularly in speech recognition systems [1,2,3]. The HMM is a kind of extension of the Markov Model. The difference is that in the HMM the current state of the model is hidden and only the output is observed (observation vector). Thus by observation of the output of the HMM, the probability of the model being in a given state can be determined. In relation to the speech recognition, the observation is the acoustic signal (in the form of an observation vector) and the state of the model is associated with the generated word (or another speech entity, such as phoneme) [4].

Classification of speech disorders includes many cases, however, their number is much lower than the number of words used in a given language. The prolonged fricative phoneme is a disturbance that often appears in a nonfluent speech. The proper detection has important significance in the further determination of the method therapy [5,6]

The recognition process with the HMM approach is as follows. First of all, it is necessary to determine the number of states of the model, as well as the size of the codebook. Next, having sufficient number of samples, a database of models can be generated – one model per one kind of a disorder. Creation of a model that recognizes a given pattern is considered to be learning. With the base model and appropriate number of encoded nonfluent utterances of the same kind, model parameters can be learned (re-estimated) so that it would be able to achieve maximum emission likelihood for that kind of pattern (observation vector). When such a database of learned models has been created, any sample can undergo examination. The recognition process consists in finding a model that

* Maria Curie-Skłodowska University in Lublin, marek.wisniewski@umcs.lublin.pl

gives the biggest probability. Since a particular dysfluency is associated with each model, that dysfluency can be detected in an acoustic signal.

The HMM is defined by three parameters $\lambda=(A,B,\pi)$, where A is the matrix of transition probabilities between particular states, B is the matrix of probabilities of emission of each element of the codebook for each model state, and π is the probability vector of the model being in a particular state at the time t=0. The problem that should be addressed is the decision about the optimal size of matrixes A and B. These values need to be chosen so that efficiency of the recognition and the time of computing is on acceptable level.

2. SAMPLE PARAMETERISATION

The acoustic signal requires to be parameterised before analysis. The most often used set of parameters in the case are Mel Frequency Cepstral Coefficients (MFCC). The process of determining MFCC parameters in the work is as follows:

- splitting signals into frames of 512 samples' length,
- FFT (Fast Fourier Transform) analysis on every frame,
- transition from linear to mel frequency scale according to the formula: $F_{mel}=2595*\log(1+F/700)$ [7, 8],
- signal frequency filtering by 20 triangular filters,
- calculation of the required (20) number of MFCC parameters.

The elements of each filter are determined by summing up the convolution results of the power spectrum with a given filter amplitude, according to the formula:

$$S_k = \sum_{j=0}^{J} P_j A_{k,j} \qquad (1)$$

where: S_k – power spectrum coefficient, J =subsequent frequency ranges from FFT analysis, P_j – average power of an input signal for j frequency , $A_{k,j}$ – k-filter coefficient.

With S_k values for each filter given, cepstrum parameter in the mel scale can be determined [9]:

$$MFCC_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n(k-0.5)\frac{\pi}{K} \right], \text{ for } n=1..N, \qquad (2)$$

where: N – required number of MFCC parameters, S_k –power spectrum coefficients, K – number of filters.

The justification of the transition from the linear scale to mel scale is that the latter reflects the human perception of sounds better.

3. CODEBOOK PREPARATION

The MFCC analysis of the acoustic signal gives too many parameters to be analysed with the application of the HMM with a discrete output. At the same time, the number of

MFCC parameters cannot be decreased, since then important information may be lost and so the effectiveness of recognition may be poor.

In order to reduce the number of parameters, encoding with a proper codebook can be applied [9]. Preparation of the codebook is as follows. First, the proper sample of an utterance needs to be chosen, which covers the entire acoustic space to be examined. Next it can be generated, for example by the use the "k-means" algorithm. Three fragments of utterances were selected, each lasting 54 seconds and articulated by three different persons and, afterwards MFCC coefficients were calculated. The obtained set of parameters were divided into appropriate number of regions and their centroids were found. For counting the distances between vectors the Euclidean formula were used:

$$d_{x,y} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \qquad (3)$$

where: $d_{x,y}$ – the Euclidean distance between "N"-dimensional vectors X and Y.

According to the described method there were several codebooks prepared with sizes 30, 38, 64, 128, 256, 512 and used in further tests.

4. TESTING PROCEDURE

The examination process was as follows. First, there were sufficient number of prolonged fricative phonemes chosen (^, s, z, x, ¥, v, •, f). For the every phoneme there were 5 fragments prepared, that contained only the prolongation. The fragments came from different recordings of stuttered people. Every group of fragments were encoded with the earlier prepared codebooks. As the result there were training vectors acquired for following codebook sizes: 30, 38, 64, 128, 256 and 512. These vectors were used for training recognition models. In the tests several models were used with sizes: 5, 8, 10 and 15 states. It means that for every prolongation there were 24 models prepared (the total number of used models was 192). As base models served models with randomly generated probability values for matrixes A, B, π.

For testing, the application named HMM was used, where appropriate algorithms were implemented. Parameters of the sound samples which were used were as follows: sample frequency: 22050Hz, amplitude resolution: 16 bits. All the records were normalized to the same dynamic range – 50dB.

The examination of the recognition effectiveness of fricative phonemes was carried on 22 fragments of utterances, each lasting several seconds. Every utterance contained only the one disorder. The test piece was encoded with every prepared codebook and then analysed by appropriate groups of learned models. From the sample, segments of the length of 10 symbols were taken (which corresponds to approximately 232 ms length) with the step of one-symbol length (approximately 23 ms) and the emission probability for each model were counted.

As the result of the recognition process the probability distribution across the time were acquired for each model. Then the time, when the maximum likelihood was revealed (achieved by whatever model), were compared with the time of the disorder appearance

(read out from the sample spectrogram). If the both were compliant it was considered as a successful recognition, otherwise as a failure.

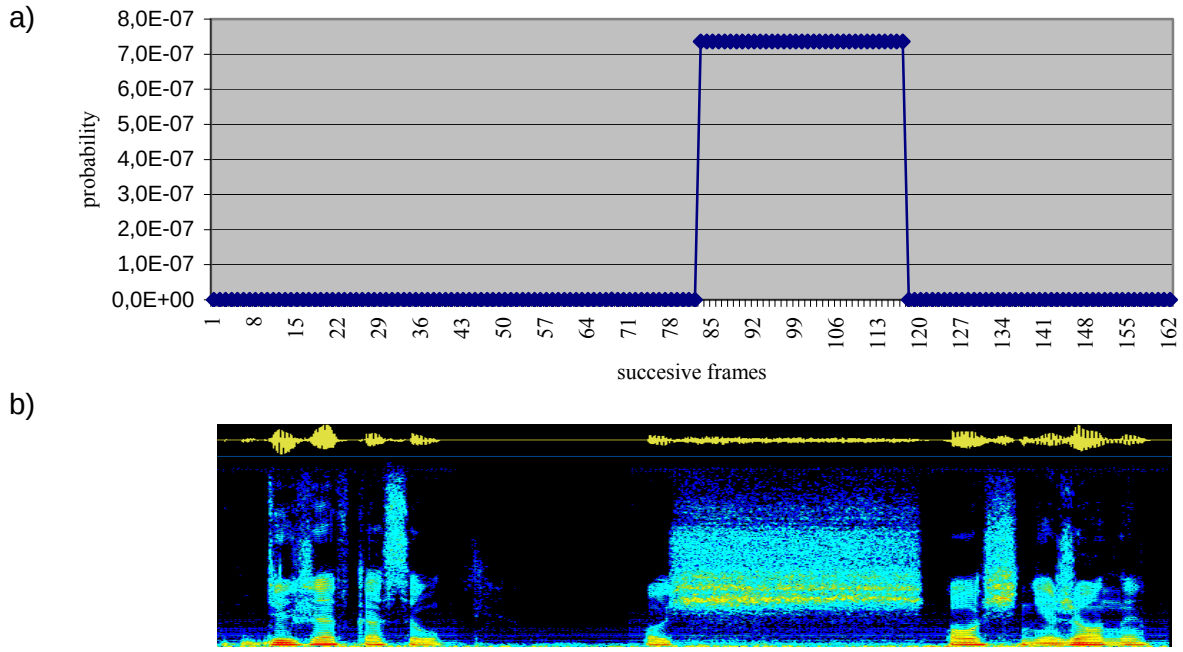

Fig.1. The analysis result of the utterance „drzewka są ośśśśś ośnieżone” (“d¥efka soN o^^^^^^^^^^ o^øe¥one”) : a) probability distribution for one of 8-state model with the codebook size of 256-elements; b) spectrogram [11].

In the figure 1a the probability distribution is shown for one of the model for the utterance „drzewka są ośśśśś ośnieżone” (“d¥efka soN o^^^^^^^^^^ o^øe¥one”).

This is the example of a very well recognition. Additionally from the graph one can estimate the duration of the disorder.

Table 1. The dependency of the recognition efficiency on the model size and the codebook size.

Number of HMM states	Codebook size	Recognition ratio [%]
5	30	64
	38	64
	64	73
	128	64
	256	77
	512	82
8	30	45
	38	59
	64	59
	128	68
	256	68
	512	77

Number of HMM states	Codebook size	Recognition ratio [%]
10	30	59
	38	45
	64	64
	128	68
	256	77
	512	82
15	30	50
	38	68
	64	73
	128	73
	256	73
	512	77

Form the table 1 it comes out that the best result, approximately 80%, were achieved for the codebook with the largest size: 512 elements, without regard for the number of states of the used model. Alongside decreasing the codebook size, the recognition ratio was also

decreasing. The influence the number of states on the recognition has the minimal importance.

5. SUMMARY

In the polish language one can distinguish 37 phonemes [12]. It seems that this number could be appropriate as a size of the codebook. But there exists a large group of sounds considered as inter-phoneme transitions. The prolongation of fricative phonemes are characterized by inclusion many sounds that are difficult to classify as a real phoneme, so the results are better for larger codebooks.

BIBLIOGRAPHY

[1] http://cmusphinx.sourceforge.net
[2] http://htk.eng.cam.ac.uk/
[3] http://julius.sourceforge.jp
[4] DELLER J. R., HANSEN J. H. L., PROAKIS J. G., Discrete-Time Processing of Speech Signals, IEEE, New York 2000.
[5] KUNISZYK-JÓŹKOWIAK W., SMOŁKA E., SUSZYŃSKI W., Akustyczna analiza niepłynności w wypowiedziach osób jąkających się, Technologia mowy i języka. Poznań 2001.
[6] SUSZYŃSKI W., Komputerowa analiza i rozpoznawanie niepłynności mowy, rozprawa doktorska, Gliwice 2005.
[7] WAHAB A., SEE NG G., DICKIYANTO, R., Speaker Verification System Based on Human Auditory and Fuzzy Neural Network System, Neurocomputing Manuscript Draft, Singapore.
[8] PICONE J.W., Signal modeling techniques in speech recognition, Proceedings of the IEEE, 1993, 81(9): 1215-1247.
[9] SCHROEDER, M.R., Recognition of complex acoustic signals, Life Science Research Report, T.H. Bullock, Ed., (Abakon Verlag, Berlin) vol. 55, pp. 323-328, 1977.
[10] TADEUSIEWICZ R., Sygnał mowy, Warszawa 1988.
[11] HORNE R. S., Spectrogram for Windows, ver. 3.2.1
[12] BASZTURA CZ., Źródła, sygnały i obrazy akustyczne, WKŁ, Warszawa 1988.