Annales UMCS Informatica AI 7 (2007) 91-100

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Automatic detection of speech disorders with the use of Hidden Markov Model

Marek Wiśniewski*, Wiesława Kuniszyk-Jóźkowiak, Elżbieta Smołka, Waldemar Suszyński

Institute of Computer Science, Maria Curie-Skłodowska University, Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland

Abstract

The most frequently used methods of automatic detection and classification of speech disorders are based on experimental determination of specific distinctive features for a given kind of disorder, and working out a suitable algorithm that finds such a disorder in the acoustic signal. For example, for detection of prolonged phonemes, analysis of the duration of articulation is used, and on the contrary, phoneme repetition can be detected with the spectrum correlation methods. Additionally, in the case of prolonged phonemes, classification based on their kind is required (nasal or whispered phonemes, vowels, consonants, etc). Therefore, for every kind of a disorder, a separate algorithm needs to be worked out.

Another, more flexible approach is the application of the Hidden Markov Models (HMM). For the needs of the presented work, the HMM procedures were implemented and some basic tests of speech disorder detection were conducted.

1. Introduction

The HMMs are stochastic models which are widely used for recognition of various patterns. They gained significance particularly in speech recognition systems [1,2,3]. The HMM is a kind of extension of Markov Models. 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 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].

The great advantage of the HMM is the simplicity of its adaptation to recognition of varying patterns coming from varying signals, as well as the

*Corresponding author: *e-mail address*: marek.wisniewski@umcs.lublin.pl

possibility of using their parameters together (audio-video speech recognition) [5].

Classification of speech disorders includes many cases, however, their number is much lower than that of words used in a given language. On the other hand, such disorders have various nature, so there is a necessity to implement separate algorithm for every kind of disorder [6,7]. Therefore, it seems that the HMMs are the ideal solution for automatic detection and classification of speech disorders.

2. Working with HMM

The recognition process with the HMM approach is shown in Figure 1. First of all, it is necessary to determine the kind of HMM model in an experimental way. The most significant thing to do here is to determine the number of states of the models, as well as the size of the codebook. Next, when the number of samples is sufficient, 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 the 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 the 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 gives the greatest probability. Since a particular dysfluency is associated with each model, that dysfluency can be detected in an acoustic signal.

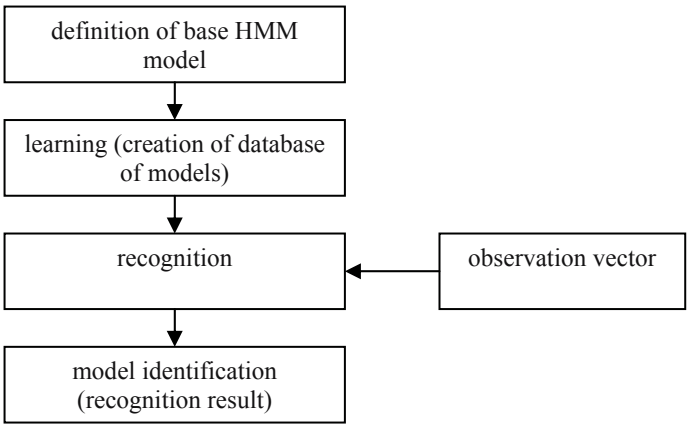

Fig. 1. The HMM based recognition process

3. Sample parametrization

The acoustic signal requires to be parameterized before the analysis. The most often used set of parameters in this case are Mel Frequency Cepstral Coefficients (MFCC). The process of determining the 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 the linear to the mel frequency scale according to the formula: F_{mel}=2595*log(1+F/700) [8,9],
- signal frequency filtering by 20 triangular filters (Figure 3),
- calculation of the required (20) number of MFCC parameters.

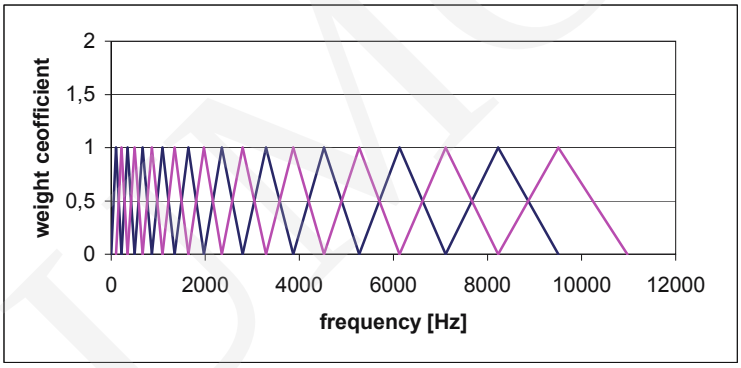

Fig. 2. The characteristics of the applied filter set (20 filters covering 0-10969 Hz range)

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:

$$S_k = \sum_{j=0}^{J} P_j A_{k,j} ,$$

where: S_k – power spectrum coefficient, J = subsequent frequency ranges from the 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 [10]:

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

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 the mel scale is that the latter reflects the human perception of sounds better.

3.1. Codebook preparation

The MFCC analysis of the acoustic signal gives too many parameters to be analyzed 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 [11]. 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. In the case of the Polish language there are 37 phonemes [12], which can be considered to be its acoustic space. However, it is not entirely true, as a lot of other sounds exist, particularly inter-phonemes transitions. Therefore, the selection of the size of the codebook needs to be done in an experimental way. In the work, a 38-element codebook was chosen (37 phonemes plus one element for the silence description). When the size of the codebook is known, it can be generated. The authors of the present work used the "k-means" to generate it. Fragments of utterances were selected (each lasting 54 seconds and articulated by three different persons) and the MFCC coefficients were calculated. From the obtained set of parameters, first 38 were chosen and recognized as initial centroids. Next, for each MFCC vector of parameters, the distances from all the vectors to the selected centroids were calculated according to the formula:

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

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

When the new distances have been calculated, the MFFC vector is assigned to the closest centroid. After all the MFCC vectors have been assigned, new positions of centroids are calcutated. Then all the MFCC vectors are reassigned to the new closest centroids. The process continues until centroids do not change positions (or changes are smaller than the defined threshold).

The obtained codebook, used for signal encoding in the work, is shown in Figure 3.

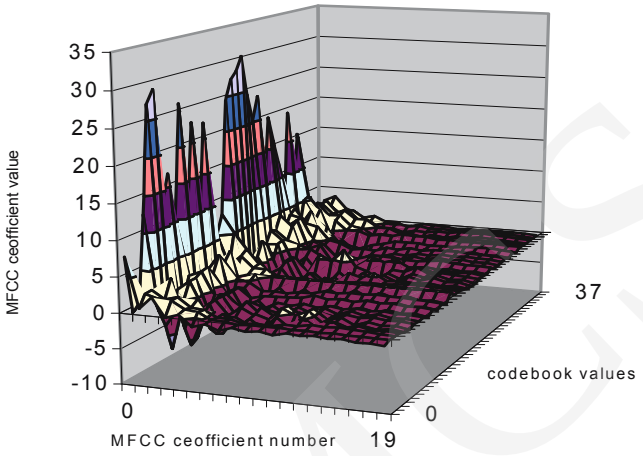

Fig. 3. The codebook used in the article – effect of the "k-means" clasterization with 38 areas

3.2. Data coding

Below the spectrograms of nonfluent utterances [13] and their equivalents after encoding with the codebook are presented. Figure 4b presents an utterance where a prolonged "s" phoneme occurs and Figure 4a shows the same fragment after encoding. The similarity of the two is clearly visible, particularly in the place where the prolonged "s" appears.

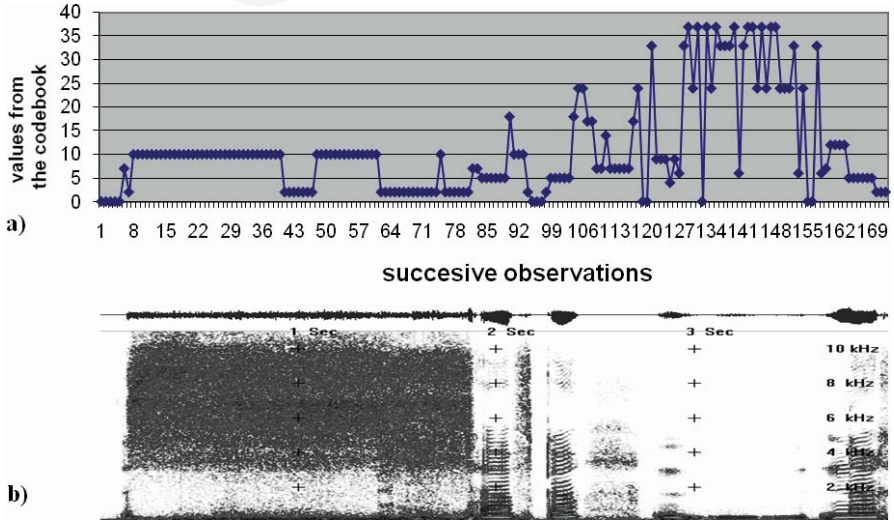

Fig. 4. The „sssssssssssawka yn yn nach" utterance – a) encoded, b) spectrogram

Figures 5a) and 5b) show an utterance including a repetition of a "k" phoneme as a spectrogram and after encoding. Also in this case coincidence is visible between the two figures, particularly during the nonfluent "k" articulation.

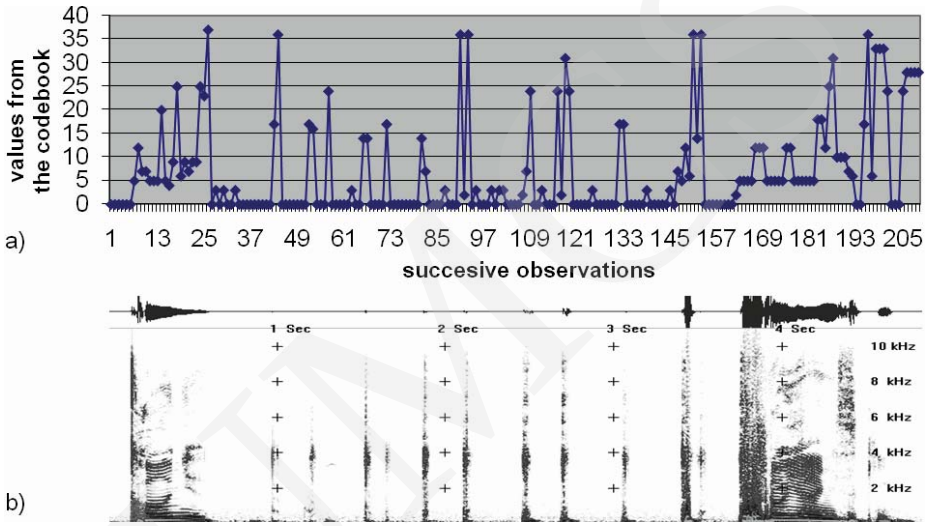

Fig. 5. The „ koń t t t t trojański" utterance – a) encoded, b) spectrogram

4. Hidden Markov Model – number of states

The HMM is defined by three parameters $\lambda = (A,B,\pi)$, where A is the matrix of probabilities of transitions 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 $t = 0$.

In the present work, the 8-state model and the codebook consisting of 38 elements were chosen for testing. In that case the matrices had the following sizes: the A matrix: 8x8, the B matrix: 8x38, and π is a vector of 8 elements. The matrices must be normalized so that the sum of all the elements of the vector π is equal to one, the sums of the row elements of matrices A and B must also be equal to one.

The problem that accompanies the choice of the proper HMM model is the number of states and the kind of emission. The 8-state model was selected on the basis of the tests described in paper [5]. The author tested the recognition efficiency of 50 isolated words for a varying number of states and for varying sizes of codebooks (Table 1). It can be observed that the best results were achieved for the 8-state model and for the codebook consisting of 37 elements. It

seems that such parameters are ideal from the point of view of efficiency as well as the required calculation time.

Table 1. The dependence of recognition efficiency on the model and the codebook sizes [5]

Number of HMM states	Codebook size	Recognition ratio [%]
5	32	88
	37	92
	64	88
	128	86
	256	84
8	32	96
	37	100
	64	98
	128	92
	256	86
10	32	86
	37	94
	64	86
	128	86
	256	84
15	32	86
	37	90
	64	84
	128	82
	256	82

5. Sample recognition with the HMM program

For testing, an application named HMM was written, where appropriate algorithms were implemented (Fig. 6). The 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. Five kinds of speech disorders were selected for examination: prolongations of "ś", "s", "z" and repetitions of "k", "t". For teaching a given model several samples containing the same kind of speech disorder were used, each sample coming from a different person. Thus the model for recognition of "k" repetition was trained with 4 samples, for "ś" prolongation – 5 samples were used, for "s" prolongation – 4 samples, for "t" repetition – 6 samples, and for "z" prolongation – 5 samples. The teaching data were normalized and encoded with the application of a 38-symbol codebook. The base model contained randomly selected probabilities for A, B, π matrices.

The recognition process was as follows. The test piece was encoded with the previously obtained codebook. Then, the segments of the length of 10 symbols were taken from the sample (which corresponds to approximately 232ms length) with the step of one-symbol length (approximately 23 ms).

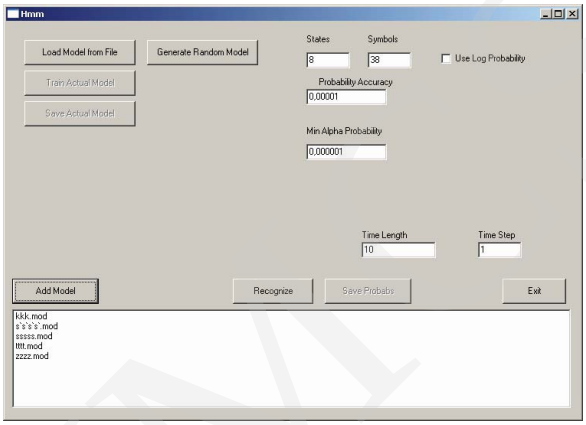

Fig. 6. The main window of the HMM program

For each segment and for every learned model the likelihood of emission was calculated. The HMM program is limited to tracking only one probability with the greatest value for a given disorder, so the examined pieces included only one kind of disorder. The recognition results are shown below.

Table 2. The input utterance: „koń t t t t trojański”

Model Name	Probability	
kkk.mod	6.05037e-29	
s`s`s`s`.mod	2.14116e-35	
sssss.mod	2.43592e-34	
tttt.mod	0.00200928	true
zzzz.mod	6.88536e-21	

Table 3. The input utterance: „na s s s s s s s s ssłoń na ssłońcu”

Model name	Probability	
kkk.mod	6.56601e-29	
s`s`s`s`.mod	1.50835e-19	
sssss.mod	0.00334747	true
tttt.mod	5.18061e-07	
zzzz.mod	0.00157918	

Table 4. The input utterance: „morza wznosiło ś ś ś ś ś się piękne miasto

Model name	Probability	
kkk.mod	8.99973e-32	
s`s`s`s`.mod	0.0699363	true
sssss.mod	1.04039e-05	
tttt.mod	0.0213177	
zzzz.mod	2.97486e-16	

Table 5. The input utterance: „w mieście k k k k Koryncie”

Model name	Probability	
kkk.mod	1.14755e-30	
s`s`s`s`.mod	0.00121974	false
sssss.mod	5.7753e-15	
tttt.mod	2.59207e-15	
zzzz.mod	3.55572e-18	

The presented results indicate significant recognition effectiveness, there being only one failure in the four tests which were conducted (Table 5). Instead of “k” phoneme repetition the system indicated a prolonged “ś”. It may have been caused by a limited number of training vectors for „s`s`s`s`.mod” model or an imperfect codebook.

6. Summary

Automatic diagnosis methods applied in the speech disorder diagnosis require continuous improvement. The main aim of that is to relieve therapists from arduous operations as well as to achieve diagnosis objectivity.

The application of Hidden Markov Models seems to be the obvious way to increase the effectiveness of automatic speech disorder detection. It is particularly significant that the system driven by HMM is quite easy to adapt for recognition of any given disorder. No new algorithms need to be implemented in order to do that, it is sufficient to teach a new model and add it to the system database.

Acknowledgements

Scientific work partially financed from the grant of Vice-Rector of Maria Curie-Skłodowska Univeristy.

The authors thank Natalia Fedan for language corrections.

References

[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] Kubanek M., *Metoda rozpoznawania audio-wideo mowy polskiej w oparciu o ukryte modele markowa*. Rozprawa doktorska, Częstochowa, (2005), in Polish.
[6] Kuniszyk-Jóźkowiak W., Smołka E., Suszyński W., *Akustyczna analiza niepłynności w wypowiedziach osób jakających się*. Technologia mowy i języka, Poznań, (2001), in Polish.
[7] Suszyński W., *Komputerowa analiza i rozpoznawanie niepłynności mowy*. Rozprawa doktorska, Gliwice, (2005), in Polish.
[8] Wahab, A., See Ng, G., Dickiyanto, R., *Speaker Verification System Based on Human Auditory and Fuzzy Neural Network System*. Neurocomputing Manuscript Draft, Singapore.
[9] Picone, J.W., *Signal modeling techniques in speech recognition*. Proceedings of the IEEE, 1993, 81(9) (1993) 1215.
[10] Schroeder, M.R., *Recognition of complex acoustic signals*. Life Science Research Report, T.H. Bullock, Ed., Abakon Verlag, Berlin, 55 (1977) 323.
[11] Tadeusiewicz R., *Sygnał mowy*. Warszawa, (1988), in Polish.
[12] Basztura Cz., *Źródła, sygnały i obrazy akustyczne*. WKŁ, Warszawa, (1988), in Polish.
[13] Horne R.S., Spectrogram for Windows, ver. 3.2.1.