Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach

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

Institute of Computer Science, Maria Curie-Sklodowska University, Pl. Marii Curie-Sklodowskiej 1, 20-031 Lublin, Poland
marek.wisniewski@umcs.lublin.pl

Summary. Hidden Markov Models are widely used for recognition of any patterns appearing in an input signal. In the work HMM's were used to recognize two kind of speech disorders in an acoustic signal: prolongation of fricative phonemes and blockades with repetition of stop phonemes.

In the work a tests results of a recognition effectiveness are presented for considered speech disorders by HMM models in different configurations. There were summary models applied for a class of disorder recognition, as well as models related to disturbance of individual phoneme. The tests were carried out by use of the author's implementation of HMM procedures.

1 Introduction

The classification of speech disorders involves many types of disturbances. Proper recognition of these has a very important significance for the choice of a therapy process [1, 2]. The equally essential problem is an objective evaluation of the kind of disorder as well as release a therapist from arduous rehearing and analyzing recorded utterances of stuttering people. So a further search for more accurate methods of automatic disturbance detection is desirable.

The HMMs are stochastic models that are widely used for recognition of various patterns appearing in an input signal. HMMs are used for description of a system state. However the state cannot be explicitly determined because it is hidden. Only observations generated by the model are given, and it is only the base on that one can estimate probability of being a system in a particular state. In the case of speech recognition systems an observation is an acoustic signal and the state is the recognized pattern (i.e. disfluency)[3].

In the recognition process with the HMM's there is a creation of the database of models required. Every model is designed for recognition of particular pattern (disfluency) appearing in a signal. Next there is a probability of emission of the analyzed fragment counted for every model in the database. The model that gives the greatest probability is then selected and if the probability is above the chosen threshold the recognition is done.

M. Kurzynski et al. (Eds.): Computer Recognition Systems 2, ASC 45, pp. 445–453, 2007.
springerlink.com © Springer-Verlag Berlin Heidelberg 2007

Every recognition model is prepared by training. Having a base HMM model $\lambda = (\pi, A, B)$ and a sufficient number of samples of the same pattern, one can prepare a model, so that it achieves maximum emission probability for that pattern.

For requirements of this paper there were several models learned designed to recognition of prolonged fricative phonemes and blockades with repetition of stop phonemes. That disturbances are the most often presented in nonfluent speech.

2 Sample Parameterization

The acoustic signal requires to be parameterized 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: $F_{mel} = 2595log(1 + F/700)$ [4, 5],
- 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:

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

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 [6]:

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

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

Division of an audio sample into frames leads to disturbance of real signal parameters. For avoidance of this problem each value of an audio frame were multiplied by proper ceofficient of the Hamming window [7], according to the formula:

$$W[n] = 0.54 - 0.56 \cos\left(\frac{2\pi k}{n-1}\right),$$

where: n- sample number, N- frame width.

However use of that formula cause losing of the information on a frame boundaries. That's way there were partial overlapping of successive frames applied. The length of an overlapp was adjusted to 1/3 of the frame length (rounded to 170 points).

3 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 [7]. 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:

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

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

For the examination there is to determine a size of the codebook. The size should be such so a recognition ratio is on an acceptable level and a computation time is reasonable. Based on own studies there were a codebook prepared with 512 elements and used for testing.

4 Testing Procedure

For tests purpose there was an audio sample of a length of 87624 ms prepared. The sample contained 24 disfluencies (10 stops blockades with repetitions of and 14 prolongations of fricatives). Every kind of disturbance appeared two times in the sample and they were: C, s, z, x, Z, v, S (fricatives) and p, t, k, b, g (stops). For every kind of disfluency there were suitable model prepared and additionally two summary models for a disturbance class recognition (stop blockades and fricatives prolongation). During the tests it has appeared that, in the case of stop blockades, a silence have negative influence on recognition ratio. To eliminate of this an additional groups of models for stop blockades were trained with the use of samples freed from silence.

Base models had 8 states and 512 code symbols. Probability values for matrixes A, B, π were randomly generated. From 3 to 6 patterns of the same

disfluency were utilized for training every model. In the case of summary models the number of patterns was much greater: for prolongation of fricatives recognition model there was 38 samples used, for stops blockade of recognition model - 30 samples and for summary model learned with patterns that were free of silence - 30 samples.

For testing, the HMM application 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 recordings were normalized to the same dynamic range and encoded with the use of earlier prepared codebook.

The examination of recognition effectiveness was carried on in the following way. From the sample, segments of the proper length were taken (30 code symbol - 465 ms and 60 code symbols - 930 ms) with the step of 1 symbol (about 15.51 ms) and then an emission probabilities for each model were counted.

From the obtained results there were distribution of emission probabilities across the time graph prepared for every model. Next, in the experimental way, the threshold value of a probability was chosen. A fragment that achieved the probability greater than the threshold was considered as disfluency. After cutting off lower probabilities from a graph there were only fragments that indicated disfluency. Fragments appearance time (read out from a graph) were compared with the time read out from a spectrogram. If both the times were equal and a disturbance was indicated by a proper model then it was considered as correct recognition.

In the figure 1 and 2 there are examples of probability distribution shown for three summary models and an utterance spectrogram[8]. Obtained probabilities cover a very big range so the logarithmic scale was used. The place where a disorder appears is characterized by a very high probability value in comparison to other places. For example the probability difference between two disorders that appears (stop and fricative at a time range 17500-19500 ms) is over 1E-20, so it indicates a stop phoneme blockade.

The value of probability depends on a window size - when the window is longer the probability is lower. One can notice that in the case of the 60 frames window length the graph has smaller fluctuations, but on the other side the 30 frames window length graph is more detailed.

There were two groups of test carried on in the work. In the first case (tables 1-4) there were used summary models for fricatives, summary models for stops as well as models learned for every individual phoneme separately. In the second case (tables 5-8), for stop phonemes there were applied models trained with samples with deleted silence (models for fricatives were the same). One should notice that models learned for individual phonemes was used for recognition of a class of disfluency (like summary models) and not for recognition of individual phoneme disturbances. Speech disorders are often sounds that are completely different form known phonemes and it is almost impossible to recognize nonfluency of individual phoneme. There are test results of recognition effectiveness shown below for the second groups of models (learned by samples with deleted silence).

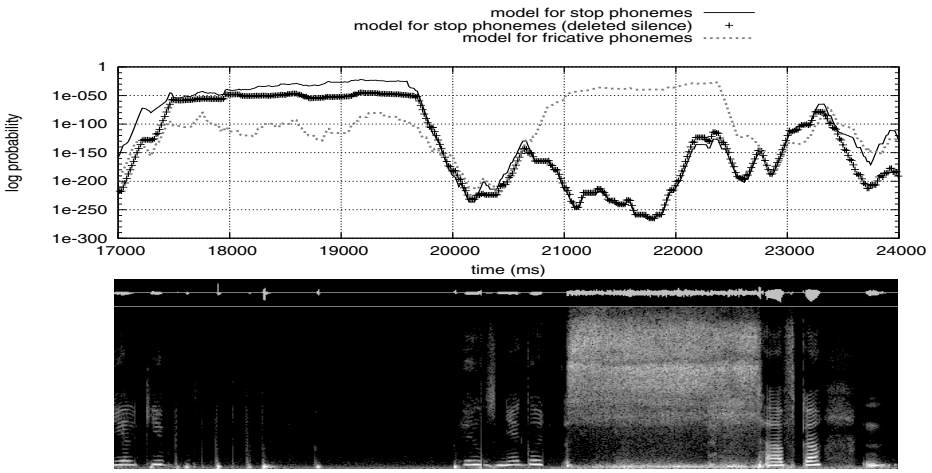

Fig. 1. The analysis result of the utterance: "vjelci bbb b1w z ego ctSwovjek ssss safka" for the window of 30 frames length (465 ms); probability distribution for three 8-state models with the codebook size of 512-elements (top); the spectrogram (bottom)

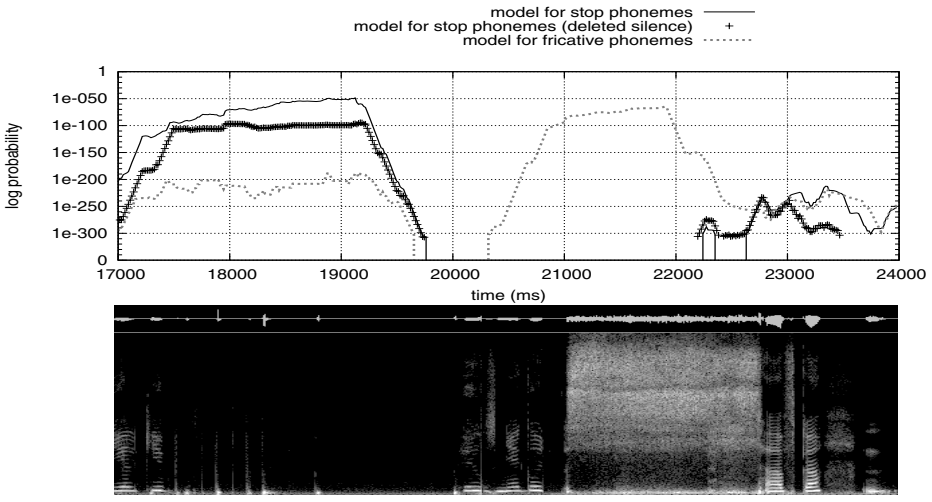

Fig. 2. The analysis result of the utterance: "vjelci bbb b1w z ego ctSwovjek ssss safka" for the window of 60 frames length (930 ms); probability distribution for three 8-state models with the codebook size of 512-elements(top); spectrogram (bottom)

For the comparison of recognition ratio two parameters are useful: sensitivity and predictability. They were counted according to formulas [9]:

sensitivity=*((number of correctly recognized nonfluencies) / (number of all nonfluencies in the sample))*100%;*

Table 1. The recognition ratio for summary models; window size 30 frames, probability threshold 1E-45

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	7	2	8	47%	80%	27%
prolongation of fricatives	0	9	5	100%	36%	36%
summary:	7	11	13	65%	54%	19%

Table 2. The recognition ratio for summary models; window size 60 frames, probability threshold 1E-90

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	0	2	8	100%	80%	80%
prolongation of fricatives	0	12	2	100%	14%	14%
summary:	0	14	10	100%	42%	42%

Table 3. The recognition ratio for models prepared for individual nonfluent phonemes; window size 30 frames, probability threshold 1E-45

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	2	4	6	75%	60%	35%
prolongation of fricatives	2	8	6	75%	43%	18%
summary:	4	12	12	75%	50%	25%

predictability=*((number of correctly recognized nonfluencies) / (number of correctly recognized nonfluencies + number of false nonfluence recognition))*100%;*

As a more general parameter the following formula was used:

correctnes =*sensitivity * (100%-predictability)*

If the above formula was giving a negative value then this value was treated as null percentage.

Table 4. The recognition ratio for models prepared for individual nonfluent phonemes; window size 60 frames, probability threshold 1E-60

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	0	2	8	100%	80%	80%
prolongation of fricatives	0	11	3	100%	21%	21%
summary:	0	13	11	100%	45%	45%

Table 5. The recognition ratio for summary models with „a deleted silence"; window size 30 frames, probability threshold 1E-49

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	2	5	5	71%	50%	21%
prolongation of fricatives	0	8	6	100%	43%	43%
summary:	2	13	11	85%	45%	30%

Table 6. The recognition ratio for summary models with „a deleted silence"; window size 60 frames, probability threshold 1E-145

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	0	2	8	100%	80%	80%
prolongation of fricatives	1	4	10	91%	71%	62%
summary:	1	6	18	95%	75%	70%

According to the results, the best recognition ratio was achieved for summary models with deleted silence (table 6) and for the window length of 60 frames (correctness equal to 70%). Such value is satisfactory. In general, the correctness was better when the window size was equal to 60 frames (only in one case was differently) and for models with deleted silence (also only in one case was differently). The predictability ceofficient was always better in every test when 60 frames window was used, so there were less incorrect recognitions.

Table 7. The recognition ratio for models prepared for individual nonfluent phonemes and with „a deleted silence"; window size 30 frames, probability threshold 1E-49

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	3	6	4	57%	40%	0%
prolongation of fricatives	3	6	8	72%	57%	29%
summary:	6	12	12	67%	50%	27%

Table 8. The recognition ratio for models prepared for individual nonfluent phonemes and with „a deleted silence"; window size 60 frames, probability threshold 1E-145

kind of disturbance	false recognitions	lack of recognitions	correct recognitions	predictability	sensitivity	correctness
stops blockades with repetition	1	4	6	86%	60%	46%
prolongation of fricatives	6	9	5	45%	36%	0%
summary:	7	13	11	67%	45%	12%

5 Summary

The recognition ratio of speech disorders with the use of HMM mainly depends on a very accurate selection of teaching patterns. In the case of blockades with repetitions the problem is a silence that appears in a sample. Sometime it leads to incorrect interpretation - silence is recognized as a disturbance. An another important conclusion is, that a proper selection of window length is crucial. The longer sections cause that there is less fluctuations and the number of incorrect recognitions is also lesser. On the other side a selection of too wide window leads to lesser sensitivity.

In the context of a recognition process very important is a selection of a proper probability threshold. It is a compromise between the sensitivity and predictability level. In the future authors plan to implement procedures for automatic selection of that threshold.

References

[1] Kuniszyk-Jóźkowiak W., Smołka E., Suszyński W.: Akustyczna analiza niepłynności w wypowiedziach osób jąkających się, Technologia mowy i języka. Poznań 2001

[2] Suszyński W.: Komputerowa analiza i rozpoznawanie niepłynności mowy, rozprawa doktorska, Gliwice 2005

[3] Deller J. R., Hansen J. H. L., Proakis J. G.: Discrete-Time Processing of Speech Signals, IEEE, New York 2000
[4] Wahab A., See Ng G., Dickiyanto, R.: Speaker Verification System Based on Human Auditory and Fuzzy Neural Network System, Neurocomputing Manuscript Draft, Singapore
[5] Picone J.W.: Signal modeling techniques in speech recognition, Proceedings of the IEEE, 1993, 81(9): 1215-1247
[6] 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
[7] Tadeusiewicz R.: Sygnał mowy, Warszawa 1988
[8] Horne R. S.: Spectrogram for Windows, ver. 3.2.1
[9] Barro S., Marin R., Fuzzy logic in medicine, Phisica-Verlag, A Springer-Verlag Company, Heidelberg, New York, 2002