Speech Defect Analysis Using Hidden Markov Models

Zdeněk CHALOUPKA, Jan UHLÍŘ

Dept. of Circuit Theory, Czech Technical University, Technická 2, 166 27 Praha, Czech Republic

chaloz1@fel.cvut.cz, uhlir@fel.cvut.cz

Abstract. *The main aim of this paper is the analysis of speech deteriorated by a very rare disease, which induce epileptic seizures in a part of brain responsible for speech production. Speech defects, represented mostly by the combination of missing and mismatched phonemes, are sought and examined in the spectral and time domain.*

An algorithm, proposed in this paper, is based on Hidden Markov Models (HMMs) and it is most suitable for the speech recognition tasks. The algorithm is able to analyze in both time and spectral domains simultaneously; in the spectral domain as a log-likelihood score and in the time domain as a forced time alignment of the HMMs.

The suggested algorithm works properly in the time domain. The results for the spectral domain are not credible, because the algorithm have to be tested on more data (not available at the time of paper preparation).

Keywords

Speech defects, LKS, developmental dysphasia, HMMs, speech recognition, forced time alignment.

1. Introduction

This paper is focused on children diseases characterized by epileptic seizures and electroencephalographic (EEG) discharges, which quite often cause the speech defects [9, 10, 11]. These diseases can be separated into the several categories according to their symptoms – Landau-Kleffner Syndrome (LKS) or acquired epileptiform aphasia, developmental dysphasia associated with epilepsy, acute aphasia (transient dysfunction of the cognitive function) [9]. In this paper we focus on speech defects caused by the developmental dysphasia, but the proposed algorithm is expected to work with the other categories as well.

The developmental dysphasia occurs to very young children, between 3 to 8 years of age. If the disease is getting cured its symptoms (speech defects) will diminish. The standard medical evaluation of treatment progression consists of a psychological and EEG investigation. During the psychological evaluation, a psychologist dictates specially chosen words (syllables, vowels, 2 to 5-syllabic words, sentences), and evaluates child's pronunciation. However, it would be convenient to have an evaluation procedure that would be independent on psychologist's subjective evaluation. Mostly, the speech defects are represented by the combination of missing, inserted or mismatched phonemes. In the previous works it was observed, that the diseased children are capable to preserve the rhythm of the word [5, 6]. But preserving the rhythm without respecting the number of word's phonemes has to affect the length of the consecutive vowels (see Fig. 1.).

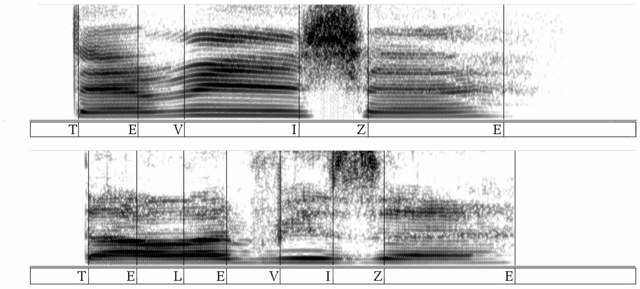

Fig. 1. Spectrogram of the word “televize” (in English “television”) uttered by a diseased (upper) and healthy (lower) child.

As it can be seen in Fig. 1, the diseased child missed the phonemes “L” and “E” (upper spectrogram), thus there is a certain spectral dissimilarity between the incorrectly and correctly uttered words. The speech defects can be easily recognized in the time domain as a distortion of length of some phonemes e.g. vowels (compare lengths of the phonemes “I” uttered by a healthy and diseased child in Fig. 1.).

There are several difficulties in the speech quality evaluation. At first, there is lack of the data available for testing. The second problem is in the age of the children. If a child is too young, then its speech defects are caused not only by the disease, but also by its still undeveloped ability to speak. Hence the algorithm was not tested (and its use cannot be suggested) with the speech data from very young children. This algorithm can be used only on the speech, in which the individual words are recognizable (the words may be mispronounced though). To obtain information about phoneme length its boundaries have to be found. The most frequently used methods are based on HMMs [12]. They allow to find the time boundaries and to measure spectral similarity (as log-likelihood) at the same time, which is quite important in this task. There are two ways to find the phoneme alignment with HMMs. The forced alignment and the recognition alignment (both realized by Viterbi algorithm [7]). The first method is more reliable, because there are no errors caused by the recognition (word error rate) [4]. The forced alignment algorithm is required to have prior information about the uttered words, because it uses phonetic transcription to obtain a HMMs sequence, which represents the uttered phonemes. The algorithm then aligns every phoneme model with a chosen section of a signal, so that the best log-likelihood score is achieved (detailed information about the forced alignment is provided in [7]).

2. Algorithm Preparation

As mentioned in the last paragraph of Section 1, the forced alignment algorithm was found to be the most suitable. This algorithm is based on HMMs, which is a statistical method, thus the HMMs have to be trained and tested (using HTK toolbox [8]) on some speech data; in this case 12 different words uttered by healthy children (only). The uttered words are the same for whole paper and can be divided into groups according to the number of the syllables – disyllabic, trisyllabic and quadsyllabic.

2.1 HMMs Training

The training data set consists of the speech of healthy children (56 speakers, both boys and girls). The data set was manually controlled to avoid speech defects caused by some logopedic disorders. Monophone HM model with three states representation (forward and loop transitions only – no skips, streams and mixtures) was used (see Fig. 2.). We also tested triphones and a monophone model with some skips (see Section 3.2). To gain the best fitted HMMs the Baum-Welch re-estimation algorithm was used repeatedly, until the convergence was achieved.

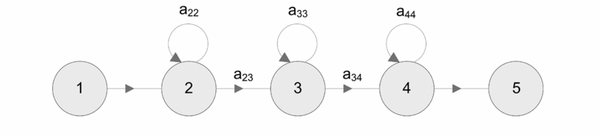

Fig. 2. The three state HM model with forward and loop transitions only (no skips) – the first and the last states are non-emitting.

2.2 HMMs Testing

The parameterization method and the used HM model have to be tested, whether they're best fitted for the mispronunciation detection. Two testing algorithms were developed. The first one measures phoneme boundary accuracy obtained by the forced alignment. The testing algorithm compares an automatic alignment to the manual alignment of the same testing data set. The manual alignment was achieved by the manual correction of the boundaries obtained by the same forced alignment algorithm. The second test is based on the properties of HMMs performing the forced alignment. In this test, the algorithm compares the log-likelihood score (obtained by the forced alignment) of the data set with a correct transcription to the log-likelihood score obtained on a data set with a transcription containing a mistake (see differences in .LAB file in Fig. 3). The differences between these two log-likelihood scores should be as high as possible. A block diagram of the second testing algorithm is shown on Figure 3.

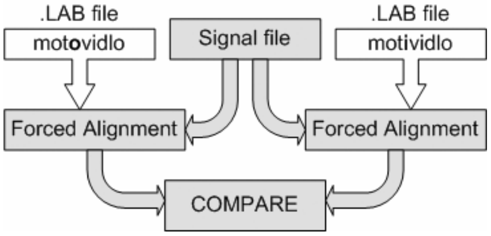

Fig. 3. The block diagram of testing algorithm.

As it can be seen from the block diagram, this testing algorithm uses the phonetic transcription of the signal files (.lab files – in HTK) to generate some mispronunciation: In this case mismatching phoneme, compare "mot**o**vidlo" versus "mot**i**vidlo". Because HTK uses monophone HM model according to the transcription in .lab file, different HMM (than the one present physically in the signal file) is used, and thus it simulates a pronounced error. In spectral domain – it should change the log-likelihood score, because the spectral parameters of the used phoneme model and the phoneme in the signal file are different. The other types of the defects (i.e. inserted or missing phoneme) can not be simulated in this manner. An inserted/missing phoneme, according to the block diagram (Fig. 3.), generates one more/less HMM (Viterbi algorithm goes through more/less states) and causes the log-likelihood score to change rapidly because of the number of states. Thus, it differs from the situation with the real data, where the number of HMMs is always the same, because the forced alignment algorithm is performed.

3. Parameter Verification

The signal parameterization and HM model suitability were tested before the real data were applied. The testing algorithms were described in Section 2.2. The optimal parameterization was determined by comparing the accuracy of the manually and automatically aligned data [4] (testing data were not included in the training set).

3.1 Parameterization Testing

The most common parameterizations were tested: LPC (Linear Prediction Coding), PLP (Perceptual Linear Prediction) coefficients, cepstral coefficients and MFCC VTLN (Mel-Frequency Cepstral Coefficients with Vocal Tract Linear Normalization). MFCC are mostly used in the ASR (Automatic Speech Recognition) systems because of their robustness to noise. This may be a convenient property as the training data were recorded in noisy environment (school). The use of the frequency scaling, such as VTLN, improves the recognition rate and thus the accuracy [1], [4]. The values in Tab. 1 represent the percentage quantity of the phonemes which boundaries (obtained with the forced alignment) differ less then 30 ms to the manual alignment. The training/testing data set consists of twenty six/sixteen healthy speakers, respectively.

LPC	PLP	VTLN MFCC
71,18	62,50	74,90

Tab. 1. Parameterization comparison.

As it can be seen, the VTLN MFCC parameterization is the most accurate. The results for the cepstral coefficients are not presented because of a low recognition rate caused by noisy training data.

3.2 HM Model Testing

Time boundary accuracy test was used to find the best HM model. Because the amount of the training data was rather limited, better results were obtained with the monophone model than with the triphone models, as the former one requires less training data. As for HM model with state skips, the second test was performed. The classic monophone model with no skips was found out to be much more accurate [4]. The streams and mixtures were tested in order to improve the mispronunciation detection. The second test, which measures the ability of detecting speech defects in the spectral domain (see Section 2.2 for more details), was performed. The data streams are formed from VTLN MFCCs and its first and second derivatives – delta VTLN MFCC and delta delta VTLN MFCC (detailed information about the streams and mixtures are provided in [7]). The next table shows the influence of the number of the chosen data streams, and the number of the mixtures on the log-likelihood score. The final value is the sum of the differences between log-likelihood score of the data set with and without transcription error (more is better).

word	1M1S	2M1S	3M1S
pivo	126.61	83.463	98.274
papír	29.733	22.809	9.6777
květina	185.4	155.3	173.7
pohádka	63.079	47.462	49.207
	1M3S	2M3S	3M3S
pivo	126.61	92.375	97.743
papír	29.733	15.383	-5.5729
květina	185.4	155.8	174.6
pohádka	63.079	41.486	36.247

Tab. 2. The sum of the difference of log-likelihood score in dependence on the number of the data streams (1S-3S) and the mixtures (1M-3M). Grey background indicates the greatest difference.

The data set with one stream and one mixture has the greatest differences (in the most cases) between the log-likelihood score of the data with and without error. The test has shown that neither the data stream nor the mixtures improve the mispronunciation detection, thus they will not be used further.

4. Main Speech Defects Testing

The forced alignment algorithm searches in the time and also in the spectral domain. At first, the HMMs' behavior has to be tested on the most frequent speech defects – a missing, inserted or mismatched phoneme. As was mentioned in Section 1, the number of HMMs (representing phonemes) is always the same (caused by phonetic transcription), but the real number of the phonemes could differ due to the wrong pronunciation. Thus the algorithm presented in Section 2.2 can not be used directly to test a missing or inserted phoneme error. In this section, the testing data were manually prepared; some of the phonemes were inserted, some cut away or mismatched. The phonemes which are often missing or mismatched (in a real situation) are very hard to evaluate, and manual modification of the signal file is very time-consuming therefore, this test was made only for a few presented words. Forced alignment was performed, and the results in the spectral and in the time domain were observed.

4.1 Spectral Domain

The log-likelihood score is the speech quality measure in the spectral domain. The forced alignment algorithm produces the log-likelihood score for every phoneme of the word. To avoid speaker’s speech rate error, the score is normalized to the sum of the segments which the HM model went through. The log-likelihood score (shown in Tab. 3.) is computed as the sum of the word's phoneme log-likelihood scores. The columns 3-5 (*+/-err*, *change*) show a shift from the original log-likelihood score (more negative is better).

Word	Log-likelihood Score			
	orig.	+ err	- err	change
papír	-345.84	-17.32	X	-13.01
dědeček	-517.97	-11.68	X	-13.12
květina	-545.3	-4.71	X	X
pohádka	-499.81	-10.57	10.46	X
pokémon	-470.57	-13.58	X	X
motovidlo	-646.08	15.8	10.97	X
popelnice	-620.56	8.08	3.54	-1.49
televize	-511.55	-2.98	-18.69	X

Tab. 3. The log-likelihood score progression; word's score with: orig. - original word, +/-err – inserted/cut phoneme, change – mismatched phoneme, X – not realized; grey background indicates a mistake.

As it was expected, the decreasing log-likelihood value indicates some mispronunciation error. The mistakes, indicated by gray background, can be clearly explained by the forced alignment behavior. In case of the insertion, HMM tries to classify the inserted phoneme as one of the phonemes in the a-priory given transcription. As a result there will be increased number of segments HMM has to go through, which will decrease the log-likelihood. In case of the missing phoneme, HMM can not find the corresponding segments, and as a result, the number of segments HMM goes through is decreased. Hence, the log-likelihood values depend on the type of mispronunciation (its spectral characteristic) and causes algorithm to fail in some cases.

4.2 Time Domain

The number of the phonemes lengths differing from average lengths is the speech quality scale in the time domain. The lengths are obtained by the forced alignment algorithm and they are normalized to word's length to avoid error caused by different speaker speech rate. Testing data are the same as in Section 4.1. The lengths of every phoneme are compared to corresponding average lengths and the number of phonemes with higher/smaller length than the average plus/minus standard deviation is computed (see Tab. 4).

Word	Number of Differing Phonemes			
	orig.	+ err	- err	change
papír	0	2	X	1
dědeček	2	2	X	4
květina	1	1	X	X
pohádka	2	2	1	X
pokémon	0	1	X	X
motovidlo	0	2	3	X
popelnice	0	1	2	0
televize	1	1	2	X

Tab. 4. The number of the phonemes differing from average lengths; word's score with: orig. - original word, +/- err – inserted/cut phoneme, change – mismatched phoneme, X – not realized; grey background indicates a mistake.

As it can be seen from Tables 3. and 4., the time domain algorithm detected mistakes correctly, while the spectral domain algorithm failed, because of the spectral similarity of mispronounced phonemes (see Section 4.1). This test is not significant enough to make conclusions, and therefore the algorithm has to be tested on the real data set.

5. Real Data Application

The application of the algorithm on the real data is discussed in this section. At first, the spectral domain was examined. It was tested, whether the data can be divided into some clusters according to the log-likelihood score. The data are supposed to range between two separate sets that correspond to the healthy and diseased children. Sixteen healthy children (not included in training data set) and four diseased children were put into this test (see Fig. 4.). The forced alignment algorithm was performed.

According to Fig. 4, the data can be divided into some clusters. The circles marked as 4 and 5 represent the same speaker in the different phases of the disease. Realization number 5 is pronounced almost without any error, but has lower log-likelihood score which is opposite to what is requested. The log-likelihood score of the word 4 is higher, because it is much shorter (more phonemes missing thus HMMs go through less segments). Hence the data can not be simply divided into clusters according to their log-likelihood score, because of the score dispersion caused by some type of the speech defects.

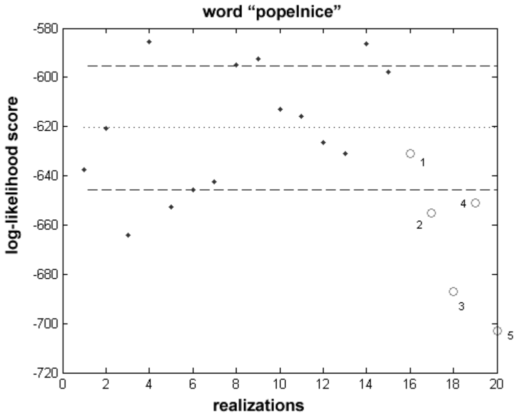

Fig. 4. Clustering testing data set; point / circle – healthy / diseased children, lines: pointed – arithmetic mean, dashed – arithmetic mean plus/minus standard deviation.

The log-likelihood score test was performed also on the time delayed speech data (about three months between two following recordings) from the same speaker. The log-likelihood score was expected to express some progression, the speech quality improving or deterioration. The quality of the pronunciation was evaluated by a listening test and the words were divided into three categories:

- 1 – good pronunciation (no mistakes),
- 2 – poor pronunciation (one to two mistakes),
- 3 – nearly unrecognizable.

The log-likelihood score compared with the listening test is shown in Tab. 5. The second column shows time delayed realizations for each word. The second and third value (for each word) is a shift from the first log-likelihood score (more positive value indicates better pronunciation).

Words	Score	List. Test
motovidlo	-747.77	3
	74.81	2
	113.72	1
popelnice	-631.03	1
	-60.31	2
	-86.29	3
sokol	-353.87	1
	-20.17	2
	-20.82	3
dědeček	-482.94	3
	-37.28	2
	-67.25	1

Tab. 5. Log-likelihood score progression; List. Test – listening test, Score – log-likelihood score; grey background - incorrectly classified.

As it can be seen, the log-likelihood increases or decreases just like the result of the listening test except the last two cases. The accuracy, tested on more data, is about 72%. This test has to be performed again when more data are available.

The test working in the time domain was performed on all data sets. The data sets consist of the sixteen children (five female and eleven male speakers) in the different phase of the disease. The quality of the pronunciation was evaluated in the same manner as in the previous paragraph. The forced alignment algorithm was performed and the lengths of the phonemes were compared to the mean lengths computed from the recordings of the healthy children (training data). Tab. 6 shows the results (only first ten speakers). The listening mark was computed as the mean of the marks of all pronounced words. The second column says how many percent of all phonemes was mispronounced.

Speaker	Wrong Phonemes	List. Test
1	46.11	2.58
2	38.64	1.84
3	24.81	1.42
4	25.97	1.33
5	40.00	1.7
6	38.46	2.27
7	46.43	2.33
8	20.78	1.42
9	50.98	2.22
10	50.00	2.81

Tab. 6. Measuring phoneme lengths; List. Test – listening test (mean of the marks), Wrong Phonemes – mispronounced phonemes (in percentage); grey background - incorrectly classified.

While the listening mark is better, the percentage of the phonemes with distorted length is lower and vice-versa. Speakers 3, 4 and 8 have less than 26% of all phonemes detected as mispronounced and their listening marks are below 1.5. Speakers 1, 6, 7, 9, 10 have more than 38% of all phonemes detected as mispronounced while their listening marks are above 2.2. Two speakers (2, 5) show difference between listening and automatic test. These speakers were unable to utter 3,4-syllabic words correctly therefore the number of wrong phonemes increased too much (4-syllabic words contain more phonemes, thus they can generate more mispronounced phonemes). If a child is able to utter only simple words, the mean of the marks will be better, while the forced alignment algorithm can detect more incorrect phonemes on more difficult words. Thus the output values (listening or obtained by algorithm) should be weighted by some coefficient in order to better define the word pronunciation difficulty.

6. Discussion

Speech defect analysis using HMMs was presented in this paper. Research was performed in the spectral and in the time domain by algorithm of the forced alignment.

In the spectral domain, it was found, that the algorithm can not be used to cluster the data into groups, because the log-likelihood changes rapidly in some cases (e.g. when the mispronounced word is much shorter to the one correctly pronounced). Research in the spectral domain could be useful while researching the speaker's speech quality progression, but the number of uttered phonemes has to be tracked in the time domain to prevent the algorithm to fail if there is a lot of phonemes missing. Still, this algorithm has to be tested on more data.

In the time domain, the algorithm was found to be more successful. The results obtained by the listening test and the forced alignment were quite the same. Different difficulty of uttered words can cause the inaccuracies. The algorithm computes output value from every phoneme and thus the 4-syllabic word (9 sounds) can produce more incorrect phonemes then 2-syllabic (4 sounds), while the listening mark has the same weight for every word. Weighting with a balancing value should correct these improper cases.

Acknowledgements

This paper was supported by the Czech Grant Agency under grant No. 102/03/H085 and by the grant MSM6840770012 Transdisciplinary Research in Biomedical Engineering II.

References

[1] STEMMER, G., HACKER, C., STEIDL, S.,NÖTH, E. Acoustic normalization of children's speech. In *EUROSPEECH-2003*, pp. 1313-1316.

[2] WARWICKER, B., LEES, J. *Landau Kleffner Syndrome*. [on-line]. February 2001. [cit. 2005-01-19]. <http://www.bobjanet.demon.co.uk/lks/home.html>.

[3] CHALOUPKA, Z., UHLÍŘ, J. Using standard algorithm and cepstral transformations for analysis of mispronunciation and incorrect phoneme sequencing. In *Digital Technologies 2004*. Žilina: Technical University of Žilina, 2004, vol. 1, p. 44-48.

[4] CHALOUPKA, Z. Analysis of mispronunciation using time alignment of phonemes. In *ESSP-2005*, p. 277-282.

[5] CHALOUPKA, Z. Mispronunciation analysing algorithm – testing methods and results. In *Digital Technologies 2005*. Žilina: Technical University of Žilina, 2005, vol. 1. ISBN 80-8070-334-5.

[6] CHALOUPKA, Z. Mispronunciation research in the time and in the spectral Domain. In *POSTER 2006 - 10th International Student Conference on Electrical Engineering* [CD-ROM]. Prague: CTU, Faculty of Electrical Engineering, 2006.

[7] YOUNG, S. *The HTKBook (for HTK Version 3.1)* [on-line]. [cit. 2005-01-19].<http://nesl.ee.ucla.edu/projects/ibadge/docs/ASR/htk/htkbook.pdf>.

[8] *HTK* [software package]. Ver. 3.2.1. Dec. 2002 [cit. 2005-01-19].

[9] SCHIRMER, C. R., FONTOURA, D. R., NUNES, M. L. Language and learning disorders. *J. Pediatr*. (Rio de J.), 2004, vol.80, no.2, suppl, p. 95-103. ISSN 0021-7557.

[10] BALLABAN-GIL, K., TUCHMAN, R. Epilepsy and epileptiform EEG: Association with autism and language disorders. *MRDD Research Reviews 2000*, vol. 6, no. 4, p.300-308. © 2000 Wiley-Liss, Inc.

[11] RAMANUJAPURAM, A. Autism and epilepsy: The complex relationship between cognition, behavior and seizure. *The Internet Journal of Neurology*, 2005, vol. 4, no. 1, ISSN 1531-295X.

[12] MEEN, D.; SVENDSEN, T.; NATVIG, J.-E. Improving phone label alignment accuracy by utilizing voicing information. In *SPECOM 2005 Proceedings*. Moscow State Linguistic University, 2005, ISBN 5-7452-0110-X, p. 683-686.

About Authors...

Zdeněk CHALOUPKA, born 1981 in Hradec Králové. He graduated at CTU in 2005 and is currently pursuing the Ph.D. degree at the Dept. of Circuit Theory, Faculty of Electrical Engineering, CTU. He specializes in speech signal processing, especially in speech recognition and synthesis.

Jan UHLÍŘ, born 1940 in Prague. He is professor at the Faculty of Electrical Engineering of Czech Technical University in Prague (CTU). He graduated at CTU in 1962 and his current research interests are analysis of electronic circuits, numerical algorithms and computer programs, digital signal processing, speech signal processing and speech recognition. He is senior IEEE member.