On the Amount of Speech Data Necessary for Successful Speaker Identification

Aleš Padrta, Vlasta Radová

Department of Cybernetics
University of West Bohemia in Pilsen, Czech Republic
apadrta@kky.zcu.cz, radova@kky.zcu.cz

Abstract

The paper deals with the dependence between the speaker identification performance and the amount of test data. Three speaker identification procedures based on hidden Markov models (HMMs) of phonemes are presented here. One, which is quite commonly used in the speaker recognition systems based on HMMs, uses the likelihood of the whole utterance for speaker identification. The other two that are proposed in this paper are based on the majority voting rule. The experiments were performed for two different situations: either both training and test data were obtained from the same channel, or they were obtained from different channels. All experiments show that the proposed speaker identification procedure based on the majority voting rule for sequences of phonemes allows us to reduce the amount of test data necessary for successful speaker identification.

1. Introduction

The term speaker recognition denotes techniques which are used for discrimination among people on the basis of their voice characteristics. There are two main groups of applications where the speaker recognition techniques can be used [1]: security applications and forensic applications. In the security applications (e.g. physical entry control, database access control, telephone transactions control), there is usually no problem with speech data, because the unknown person wishes to be recognized and therefore is willing to provide the system with such an amount of speech that is needed in order to reach a decision.

The situation is, however, quite different in the forensic applications. Here the speaker often refuses to cooperate with investigators and does not want to provide enough speech data for the speaker recognition system. An interesting question then arises, namely what is the minimal amount of data necessary for a decision about the identity of the speaker.

Whereas the minimal amount of data necessary for speaker recognition system training has been discussed e.g. in [2], we have not found any information about the minimal amount of test data in the literature. Therefore we describe our effort in this paper, the goal of which is to find the dependence between the speaker identification performance and the amount of test speech data. We suppose that the speaker identification system is based on hidden Markov models (HMMs) of phonemes. Then several identification procedures can be used. Their principles are explained in Section 2. Next, in Section 3, a detailed description of our experiments is provided, and the achieved results are discussed. The structure of speech data used for the experiments is described in Section 3 as well. A conclusion is given in Section 4.

2. Speaker identification procedures

Assume that there is a group of J reference speakers and that each speaker is represented by a set of I HMMs of phonemes. We denote the i-th model of the j-th speaker as $M_j(i)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Further suppose that a test utterance is segmented into parts S_k, $k = 1, \ldots, K$, in such a way that each segment S_k corresponds to a phoneme. K is the number of phonemes in the test utterance. Each segment S_k is marked with an index I_k which expresses the order of the phoneme represented by the segment S_k in the Czech phonetic alphabet. An illustration of this process is given in Fig. 1. For simplicity we will call the segments "phonemes" from now on.

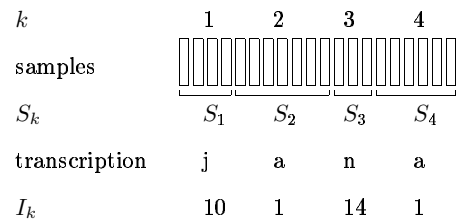

Figure 1: *An illustration of assigning indexes I_k to segments S_k.*

Suppose that our goal is to identify which of the reference speakers has spoken a given test utterance. The identification procedure is based on the Viterbi algorithm. If we know the true text of the test utterance, we can use the forced Viterbi algorithm in order to determine the log likelihood $p(S_k, M_j(I_k))$, i.e. the likelihood that the phoneme S_k is modelled by the model $M_j(I_k)$, where $M_j(I_k)$ is the model of the phoneme I_k of the speaker j. Having the likelihood $p(S_k, M_j(I_k))$ we can now use several procedures for assigning the utterances to a speaker.

2.1. Identification based on the likelihood of the whole utterance

The identification procedure often used in speaker recognition systems (e.g. [2], [3]) is based on the likelihood that the whole utterance was spoken by the speaker j. The likelihood is determined as

$$p_j = \sum_{k=1}^{K} p(S_k, M_j(I_k)). \quad (1)$$

The speaker with the highest p_j is then selected as the one who has spoken the utterance, i.e. the resultant speaker is determined

according to the formula

$$R = \arg \max_{j=1,\ldots,J} p_j = \arg \max_{j=1,\ldots,J} \sum_{k=1}^{K} p(S_k, M_j(I_k)). \quad (2)$$

2.2. Identification based on the majority voting rule for single phonemes

Since the majority voting rule proved to be a good tool for performance enhancement in our previous speaker identification experiments [4], we designed an identification procedure that is based on the majority voting rule. In that case we first identify the speaker of each phoneme S_k, $k = 1, \ldots, K$, of the test utterance according to the formula

$$R(S_k) = \arg \max_{j=1,\ldots,J} p(S_k, M_j(I_k)). \quad (3)$$

It means the speaker with the highest $p(S_k, M_j(I_k))$ is identified as the speaker who produces the phoneme S_k. We denote such a speaker $R(S_k)$. Then we compute how many phonemes of the test utterance were assigned to single speakers, and the speaker with the highest number of phonemes is identified as the one who has spoken the whole utterance. If there are two or more speakers with the highest number of phonemes, none of the reference speakers is selected as the speaker of the utterance.

2.3. Identification based on the majority voting rule for sequences of phonemes

The speaker identification procedures described in Sections 2.1 and 2.2 can be regarded as boundary cases of a general identification procedure: in one case we use all phonemes together in order to obtain a decision about the identity of the speaker, in the other case we use each phoneme separately. So let us now try to design the general procedure that will also be able to use parts larger than a phoneme but shorter than the whole utterance for speaker identification.

Denote the sequence of N successive phonemes of the test utterance which starts with the phoneme S_l and ends with the phoneme S_{l+N-1} as C_l. It means $C_l = [S_l, S_{l+1}, \ldots, S_{l+N-1}], l = 1, \ldots, K - N + 1$, where K is the number of phonemes in the test utterance. The likelihood that the sequence C_l was spoken by the speaker j is

$$p_j(C_l) = \sum_{n=0}^{N-1} p(S_{l+n}, M_j(I_{l+n})). \quad (4)$$

Now we determine the speaker of the sequence C_l according to the formula

$$R(C_l) = \arg \max_{j=1,\ldots,J} p_j(C_l), \quad (5)$$

and, then, according to the majority voting rule, the speaker to whom the greatest number of sequences from the test utterance was assigned is identified as the speaker of the whole utterance.

3. Description of experiments

3.1. Speech data

A part of the UWB_S01 corpus was used in our experiments. The UWB_S01 corpus is a read-speech corpus originally designed for training and testing speech recognition systems [5]. It consists of the speech of 100 speakers (64 male and 36 female). Each speaker read 150 sentences that were divided into 2 groups: 40 sentences were identical for all speakers, and the remaining 110 sentences were different for each speaker. The corpus was recorded in an office room where only the speaker was present. The notebook IBM TP 760 ED was used for the recording, because it has no fan and therefore its operation is very silent. However, some noise from the neighbouring offices could sometimes be heard in the recording room. Each utterance was recorded by two different microphones simultaneously. A close-talking microphone (Sennheisser HMD 410-6) recorded utterances of a high quality, whereas a desk microphone (Sennheisser ME65) recorded utterances including common office noise. Such an arrangement yielded two different recordings of each utterance. The recordings are identical in timing, but they differ in the amount of noise that they contain. Signals from both microphones were sampled at 44.1 kHz with 16-bit resolution.

All utterances of the corpus were annotated after the recording phase. The main goal of the annotation was to obtain the true text of the spoken utterances, including mispronunciation and unintelligible pronunciation. However various non-speech events and various kinds of noise were also marked. Detailed rules for annotation are given in [6].

Only the utterances of each speaker which correspond to the 40 sentences identical across all speakers were used in the experiments described in this paper. They were divided into two parts: 35 utterances of each speaker were used for training the HMMs of the speaker, the remaining 5 utterances of each speaker were used for tests.

3.2. Front-end and acoustic modelling

All utterances (both training and test) were parameterized using a 25 ms-long Hamming window with a 15 ms overlap. The dimension of each feature vector is 39 (energy and 12 mel-frequency cepstral coefficients augmented by the corresponding delta and delta-delta coefficients).

The speaker recognition system is based on continuous density HMMs of phonemes. Each phoneme is modelled by 3 left-to-right states (Fig. 2) with 2 Gaussian mixtures per state. Since the Czech language has 43 different phonemes [7], each speaker is represented by the 43 HMMs modelled from the training utterances using the HTK toolkit.

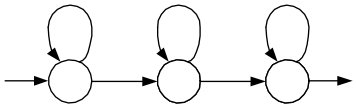

Figure 2: *Employed type of HMMs.*

3.3. Experimental results

In order to find the dependence of the speaker identification performance upon the amount of the test data, the number of phonemes K of test utterances was gradually changed from 1 to K_{max}. It means that at first only the first phoneme of each test utterance was used for speaker identification, then the first two phonemes were used, and so on. The shortest test utterance consisted of 90 phonemes, therefore we set $K_{max} = 90$. Since there were 5 test utterances for each of the 100 speakers, we could carry out 500 identification tests for each $K = 1, \ldots, K_{max}$.

First we investigated the situation when both the training and test data originated from the close-talking microphone. The

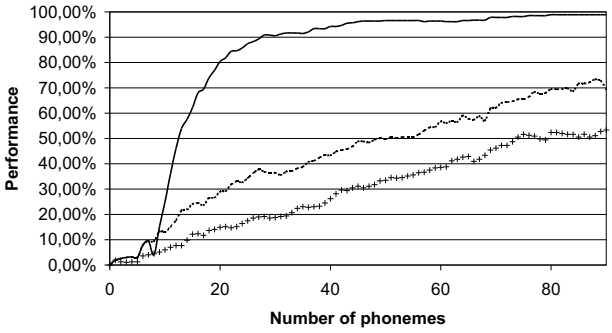

Figure 3: *Speaker identification performance when both the training and test data were obtained from the close-talking microphone. Dashed line = identification based on the likelihood of the whole utterance; line of crosses = identification based on the majority voting rule for single phonemes; solid line = identification based on the majority voting rule for sequences of 7 phonemes.*

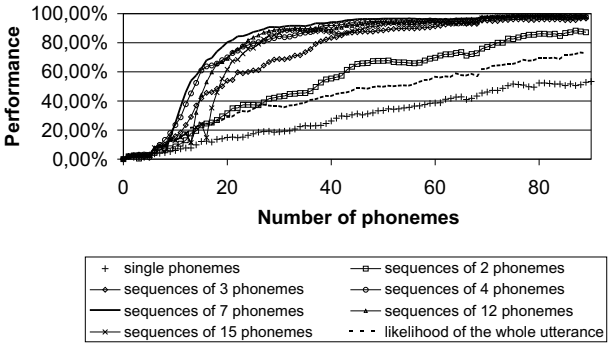

Figure 4: *Determination of the optimal number of phonemes for the procedure based on the majority voting rule for sequences of phonemes.*

Table 1: *Number of phonemes necessary for achieving the specified speaker identification performance using the procedure based on the majority voting rule for the sequences of N phonemes.*

N	50%	75%	90%	95%	98%
2	38	69	–	–	–
3	19	36	57	69	–
4	14	23	38	52	77
5	13	18	34	43	72
6	13	18	28	44	73
7	13	19	28	43	73
8	13	20	29	43	77
9	14	20	31	48	71
10	15	21	30	45	69
11	15	22	30	51	70
12	16	22	42	59	74
13	17	23	34	61	74
14	18	24	45	69	74
15	19	24	47	69	74
16	20	25	50	69	–
17	21	26	53	72	–

results achieved using the identification procedure based on the likelihood of the whole utterance are depicted in Fig. 3 with a dashed line. The line of crosses represents the results of the procedure based on the majority voting rule for single phonemes. As we expected, the speaker identification performance increases quite linearly with the number of phonemes used in the experiments. The procedure based on the likelihood of the whole utterances can be regarded as better than the procedure based on the majority voting rule for single phonemes, because, with the first mentioned procedure, an increase in the number of phonemes by 20 causes an increase in the speaker identification performance of about 14%, whereas with the second mentioned procedure it causes an increase of only 8%.

In order to get results for the procedure based on the majority voting rule for sequences of phonemes, we had to determine the number N of the phonemes in the sequences first. We carried out several experiments to find an optimum length of the sequences. The results of these experiments are presented partly in Fig. 4, and partly in Table 1. In the first column of the table the number N of the phonemes in the sequences is given. The next 5 columns present the minimum number of phonemes necessary for achieving the performance specified in the first row of the table using the sequences of N phonemes. A dash instead of a number in some cells of the table means that the specified performance was not achieved for the given number of phonemes in the sequences.

After an inspection of the results in Fig. 4 and in Table 1 we can say that the optimum length of sequences is 6–8, because, using such sequences, a relatively high performance can be reached very quickly (it means with a small number of phonemes). For comparison with the other two procedures we depicted the speaker identification performance for the identification procedure based on the majority voting rule for sequences of 7 phonemes in Fig. 3 with a solid line. We can see that the procedure based on the majority voting rule for sequences of 7 phonemes highly outperforms the other two procedures. In order to reach, for example, the performance of 50% it needs only 16 phonemes, whereas the procedure based on the likelihood of the whole utterances needs 49 phonemes, and the procedure based on the majority voting rule of single phonemes needs as many as 74 phonemes. In addition, the performance of the procedure based on the majority voting rule for sequences of 7 phonemes reaches 90% when it exploits 25 phonemes, but the other two procedures did not reach such a performance even when all the test data we had were used. It means that the procedure based on the majority voting rule for sequences of phonemes could be a promising way of reaching a good speaker identification performance with a small amount of test data.

In order to confirm or disprove the conclusion just presented, we carried out similar experiments also in the cases when both the training and test data came from the desk microphone, when the training data came from the close-talking microphone and test data from the desk microphone, and, eventually, when the training data were obtained from the desk microphone and the test data from the close-talking microphone. The results of these experiments are presented in Figures 5, 6, and 7, respectively. After an inspection of Fig. 5 we can come to similar conclusions as with Fig. 3. The optimum length of sequences of phonemes was also nearly the same, namely 7–9.

Figures 6 and 7 show a decrease in the speaker identification performance, which is probably caused by the different microphones used for the recording of the training and test data.

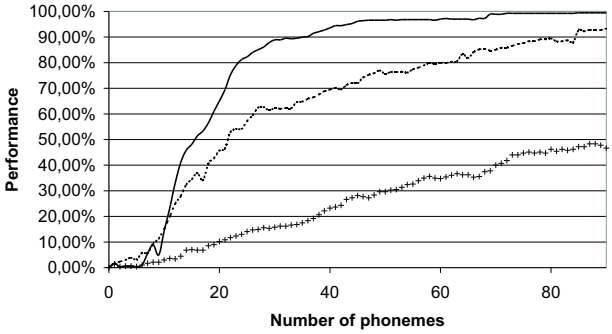

Figure 5: *Speaker identification performance when both the training and test data came from the desk microphone. Dashed line = identification based on the likelihood of the whole utterance; line of crosses = identification based on the majority voting rule for single phonemes; solid line = identification based on the majority voting rule for sequences of 8 phonemes.*

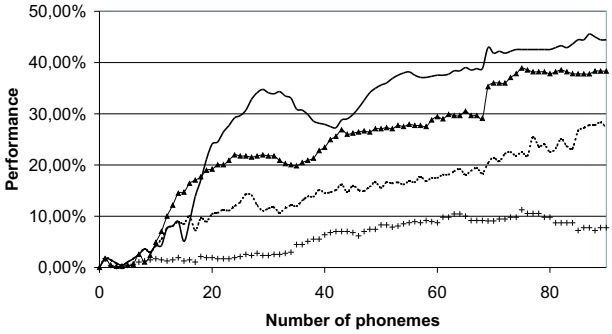

Figure 6: *Speaker identification performance when the training data came from the close-talking microphone and the test data from the desk microphone. Dashed line = identification based on the likelihood of the whole utterance; line of crosses = identification based on the majority voting rule for single phonemes; solid line = identification based on the majority voting rule for sequences of 14 phonemes; line of triangles = identification based on the majority voting rule for sequences of 7 phonemes.*

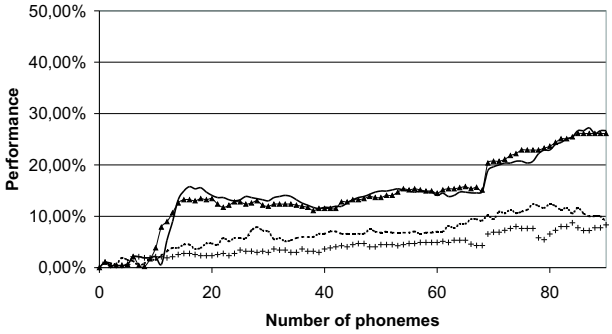

Figure 7: *Speaker identification performance when the training data came from the desk microphone and the test data came from the close-talking microphone. Dashed line = identification based on the likelihood of the whole utterance; line of crosses = identification based on the majority voting rule for single phonemes; solid line = identification based on the majority voting rule for sequences of 10 phonemes; line of triangles = identification based on the majority voting rule for sequences of 7 phonemes.*

However, the procedure based on the majority voting rule for sequences of phonemes produces still better results than the other two procedures. It is true that the optimum length of the sequences is higher than in the case when both the training and test data were recorded through the same microphone, nevertheless, also for the same length of the sequences as in the case of the same microphones, the procedure based on the majority voting rule for sequences of phonemes gives better results than the other two procedures. This fact also coincides with the conclusion which can be drawn after an inspection of Fig. 4: the identification procedure based on the majority voting rule for sequences of phonemes is always better than the other two procedures described in this paper regardless of the fact how many phonemes are used in the sequences.

4. Conclusion

The goal of this paper was to study the dependence of the speaker identification performance on the amount of test data. Three identification procedures were presented. All of them are based on the hidden Markov models of phonemes, but they differ in the way in which they deal with the phonemes of the test utterances. Several speaker identification experiments in a closed set were performed. All procedures showed quite logically that more test data cause higher performance of the speaker recognition system. However, the procedure based on the majority voting rule for sequences of phonemes, which is proposed in this paper, makes it possible to reach a relatively high speaker recognition performance very quickly. Therefore it can be regarded as a useful speaker identification procedure in cases in which the amount of test data is small.

5. Acknowledgements

The work was supported by the Grant Agency of the Czech Republic, project no. 102/02/0124, and by the Ministry of Education of the Czech Republic, project no. MSM 235200004.

6. References

[1] Doddington, G. R., “Speaker Recognition – Identifying People by their Voices”, Proc. of the IEEE, 73(11):1651–1664, 1985.

[2] Tishby, N. Z., “On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition”, IEEE Trans. on Signal Processing, 39(3):563–570, 1991.

[3] Che, C.W, Lin, Q., Yuk, D.-S., “An HMM Approach to Text-Prompted Speaker Verification”, Proc. ICASSP’96, Atlanta, USA, pp. 673–676, 1996.

[4] Radová, V., Psutka, J., “An Approach to Speaker Identification Using Multiple Classifiers”, Proc. ICASSP’97, Munich, Germany, pp. 1135–1138, 1997.

[5] Radová, V., Psutka, J., “UWB_S01 Corpus – A Czech Read-Speech Corpus”, Proc. ICSLP 2000, pp. 732–735, Beijing, China, 2000.

[6] Radová, V., Psutka, J., “Recording and Annotation of the Czech Speech Corpus”, In: Text, Speech and Dialogue, Proc. of the 3rd Workshop on Text, Speech, Dialogue, Springer-Verlag, Berlin Heidelberg, 2000, pp. 319–323.

[7] Nouza, J., Psutka, J., Uhlíř, J., “Phonetic Alphabet for Speech Recognition of Czech”, Radioengineering, 6:16–20, 1997.