Quasi Text-Independent Speaker Verification with Neural Networks

Michael Gerber and Beat Pfister

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich
{gerber,pfister}@tik.ee.ethz.ch

Abstract

A new approach to speaker verification (SV) is developed, which decides for two given speech signals, whether they have been spoken by the same person or not. The idea is to apply pattern matching for two speech signals, which are not exactly equally worded, but contain equally worded segments. This is done by first searching equally worded segments which are subsequently used for pattern matching and finally for taking the decision. For the DTW-based pattern matching, instead of an Euclidean distance measure the probability that the corresponding frames are from the same speaker is used. This probability is determined by an appropriately trained neural network. This approach to SV is completely language-independent, to some degree text-independent and it needs no speaker enrollment.

1. Introduction

In some applications it has to be decided if two speech signals have been spoken by the same person or not without a speaker enrollment. If the two speech signals are identically worded, a text-dependent method such as pattern matching can be used. If the wordings of the two speech signals are completely different, text-independent approaches such as Gaussian mixture models (GMM) are normally used.

In the case where not exactly the same was spoken, but equally worded segments are present in the two speech signals, a quasi text-independent method can be used, which determines from these equally worded segments, whether the two signals were spoken by the same speaker or not. Such a quasi text-independent approach is pursued in this work.

In section 2 the method is described and section 3 shows results obtained from both, the quasi text-independent approach presented in this work and text-dependent SV.

2. Method

The approach is to search in a first step equally worded segments in two given speech signals. In a second step the probabilities that the frame pairs along the warping curve of these phonetically matching segments were uttered by the same speaker are computed frame by frame. Finally the global probability that the two speech signals were spoken by the same speaker can be calculated from these frame-level probabilities.

2.1. Searching equally worded segments

In order to search equally worded segments, the posterior probability that two frames come from the same phoneme is calculated for all possible frame combinations. A neural network with a structure as described in section 2.3 is used to calculate this posterior probability.

These frame-level posterior probabilities are aligned in a matrix spanned by the two speech signals. This matrix is referred to as phonetic probability matrix. An example is shown in Figure 1. Phonetically similar segments show as ridges of about $45\,^\circ$ direction in the phonetic probability matrix.

Once this phonetic probability matrix is computed, contiguous sequences of frames in about $45\,^\circ$ direction with a high posterior probability have to be searched. This problem is related to the search of the optimal path in a distance matrix which is known from text-dependent SV based on pattern matching and is there tackled by dynamic time warping (DTW). The DTW algorithm is adapted to this case as follows: Instead of forcing the warping curve to start at the begin of the frame sequences of the two signals and to stop at their ends, a warping curve is started at a point of high posterior probability and follows the ridge in diagonal direction of the matrix both in forward and backward direction until the local posterior probability falls below a certain threshold indicating the end of the common segment. The segments are required to have a minimal length of about 200 ms.

The output of this algorithm is a series of pairs of phonetically similar frames.

2.2. Calculation of frame-level SV scores

A second neural network which has again the same structure as described in section 2.3 is used to compute the

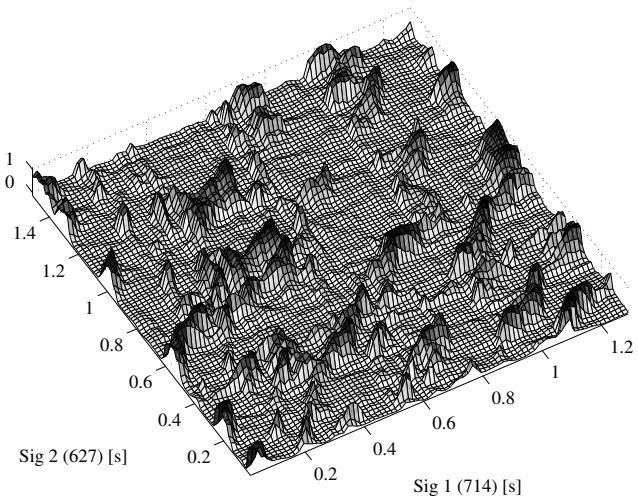

Figure 1: Probability matrix spanned by the German numbers 714 (siebenhundertvierzehn) and 627 (sechshundertsiebenundzwanzig) uttered by different speakers. For example the word hundert (hundred) which occurs in both numbers shows as a ridge between 300 and 900 ms in the first signal and between 300 and 700 ms in the second signal.

posterior probability that the two frames of a pair originate from the same speaker.

2.3. Structure and training of the neural networks

The neural networks in this work are used to calculate for two given frames the posterior probability, that they belong to the same class:

$$P(class(x_1) = class(x_2)|x_1, x_2)$$

where x_1 and x_2 are the feature vectors of the respective frames from the two signals. For the phonetic neural network this means to determine the posterior probability that the two frames are from the same phoneme. For the speaker discriminating neural network this means to calculate the posterior probability that the two frames are from speech signals of the same speaker.

A simplified structure of the used multi-layer perceptrons is shown in Figure 2. At the input of the networks the feature vectors of the frames are applied. The networks are trained to output 1 if the two frames are from the same class and 0 otherwise.

Both neural networks are speaker-independent.

3. Results and conclusions

Experiments were made with a SV database containing three-digit numbers recorded from land-line phones.

Quasi text-independent and text-dependent experiments were performed. For text-dependent experiments

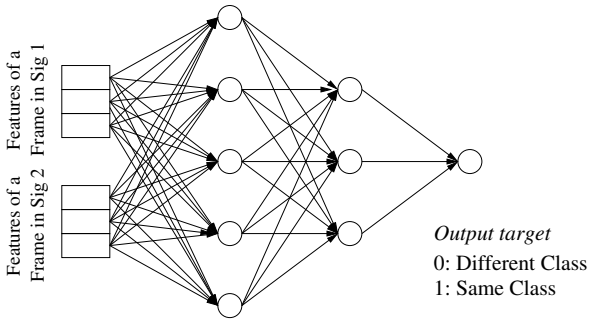

Figure 2: Simplified structure of the classification multilayer perceptron

Approach	Speaker discr.	EER (frame level)
text-dep.	Euclidean dist.	~ 0.36
text-dep.	neural network	~ 0.30
text-indep.	neural network	~ 0.35

Table 1: Equal-error ratio at frame level for text-dependent and text-independent setups

signals with the same number were taken whereas for quasi text-independent experiments only signals with different numbers were used. In the quasi text-independent case only frames from segments longer than 200 ms were used to calculate the probability that the signals are from the same speaker. The resulting equal error ratios (EER) at frame level are shown in Table 1.

For the text-independent case an average of about 0.25 seconds of matching segments was found in two three-digit number utterances which are on average about 1.5 seconds long. It was observed that the frame-level EER for the text-independent experiments using a neural network for speaker discrimination was even lower than the EER for the text-dependent experiments which use Euclidean distance for speaker-discrimination. Yet for the text-independent case the EER at frame level was still higher than for the text-dependent case, if a neural network was used for speaker-discrimination. One reason for this may be that in text-dependent SV speaker-specific pronunciation variants help to discriminate between speakers, whereas in the new approach they are excluded. Furthermore, the search for identically worded segments is not completely correct.

4. Outlook

An issue that demands for further investigations is how to compute from the frame level SV scores (i.e. the posterior probabilities that frame pairs are from the same speaker) a global one which gives the maximum correct decision rate.