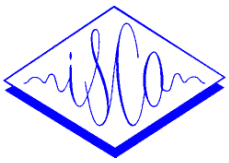

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

# SEPARATION OF SPEAKERS IN AUDIO DATA

Jesper Ø. Olsen

Center for PersonKommunikation, Aalborg University,
Fredrik Bajers Vej 7, DK-9220 Aalborg Øst, Denmark

email: jo@cpk.auc.dk

## ABSTRACT

Speaker separation is a technique with potentially many applications, for instance as an aid in browsing audio documents. This paper describes a novel speaker separation method, where speaker models are created without having any training data available in advance. The method was tested on realistic unconstrained telephone conversations, and ergodic Hidden Markov Models used for speaker modelling. The overall results were sequence and duration accuracies of respectively 87% and 94%, when no prior knowledge of the speakers was used (i.e. training data).

**Keywords:** Speaker Separation, Speaker Recognition, Hidden Markov Models.



## 1 INTRODUCTION

Speech can easily be recorded in "audio documents", but compared to text documents access is more difficult: Listening is slower than reading and audio can not easily be searched for keywords. Often it is desirable to access certain parts of a recording, rather that listening to it sequentially from end to end, but because of the apparent lack of structure, this kind of browsing can be a difficult and tedious task. Since many workstations today are equipped with audio I/O facilities, it is an obvious choice to use speech processing to provide audio documents with some kind of structure. Although not many tools have been developed for this, several research projects have investigated possible solutions [1]. For many purposes speaker separation and identification is a key component for processing audio recordings of conversations between two or more speakers. Possible applications of this include:

- Visualisation of audio data as an aid for browsing and manipulating (moving, copying and deleting) speech segments.
- Transcription of a conversation; who said what.
- Extraction of one side of a conversation for further processing.

In this study, speaker separation algorithms for the limited case of conversations between two speakers were investigated. Three specific task constraints were considered:

1. Both speakers known. The ideal case, not always realistic, but useful as a baseline result.
2. One speaker known. In many applications a "user" model can reasonably be assumed available, for instance when the person who recorded the audio is also one of the speakers.
3. Neither speaker known. The general case.

The general approach was to use Hidden Markov Models (HMMs) as speaker models, and for this purpose the HTK Toolkit [2] was used.

## 2 METHOD

Ergodic HMMs can be used for speaker modelling in text independent speaker recognition [3, 4]. It is thought that the individual states in the models will model phones or phone classes of a speaker. If an ergodic HMM (see figure 1) is trained on the speech of not one, but two or more speakers, then it is intuitively plausible that – if the model has enough states – the individual states will model phones of just one of the two speakers. If only one of the speakers is speaking at any one time, transitions between intra-speaker states (phones) will be much more common than transitions between inter-speaker states (phones). Consequently the transition probabilities between "same speaker" states can be expected to be much higher than "cross speaker" transition probabilities. This fact (hypothesis) can be exploited by connecting two ergodic HMMs in parallel (a 2xERG type HMM, see figure 2). When this model is trained on speech from two speakers, then the states belonging to different speakers will cluster in different parts of the compound model, because only thereby will a

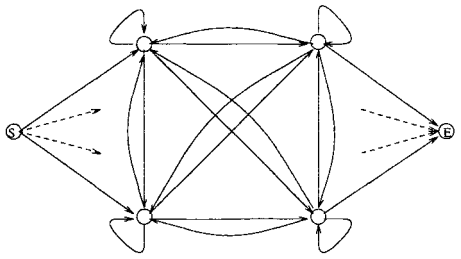

Figure 1: A 4 state ergodic HMM; a ERG4 HMM. Well suited for text independent speaker modelling.

maximum (sufficient) number of intra-speaker transitions be available. After training, the two sides of the 2xERG type model can easily be disconnected, and two fully trained speaker models thereby be made available.

## 2.1 The Speech Data

The speech data used in this study were four telephone conversations from the SWITCHBOARD corpus [5]. This corpus has been supplied with a set of word level time-aligned transcriptions. Silence periods have not been transcribed, and can not be directly inferred from these files, which will be referred to here as the reference segmentation. The conversations were between different combinations of male and female speakers. None of the speakers took part in more than one conversation, so each conversation (average length 5 min) had to be treated separately. The speech data was parameterised as 32 dimensional vectors (using a 25.6 ms Hamming window and a frame period of 10 ms): 15 Mel Frequency Cepstral Coefficients (MFCCs) + normalised log energy + delta MFCCs + delta log energy.

## 2.2 Experiments

A set of experiments corresponding to the three task constraints were conducted. In each of these, three different HMMs were trained: two ergodic speaker models (A and B) and a 4-state left-to-right silence model (SIL). The ergodic models had three Gaussian mixtures per state, whereas the silence model only had one mixture per state. When trained, a recognition network (syntax) was set up with the three models connected in a parallel loop, and speaker separation performed by Viterbi decoding the speech data. As will be described in section 2.3, this segmentation was subsequently subjected to a post processing step in which a more "coarse grain" segmentation was computed by deletion of some segment boundaries and reclassification of the new compound segments.

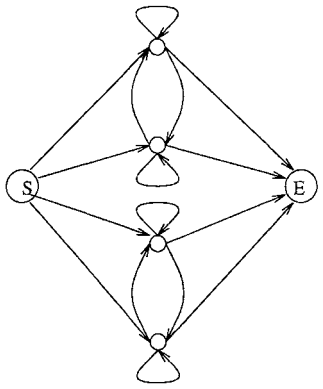


Figure 2: A parallel coupling of two ergodic HMMs; a 2xERG2 HMM. When trained on speech from two speakers, the two halves of the model will correspond to different speakers.

*2.2.1 Experiment 1, Baseline*

The separation was done using ergodic speaker models with in turn 1, 3, 6, 12 and 24 states. The reference segmentation was used as the basis for maximum likelihood training of these models. Ideally training and test data should have been different, but as there was only around 100 seconds (a third of a conversation) of speech available for each speaker, this ideal was disregarded.

*2.2.2 Experiment 2, User Dependent*

In this experiment, a pretrained speaker A model was used (the 24 state model from experiment 1). The following was done in order to create a model for speaker B and silence:

1. By examining the signal energy, speech was separated from silence. This was in order to avoid training the speaker models on silence periods, and to provide a segmentation into segments that with a high probability had been uttered by just one speaker. A silence period does not in general suggest a change of speaker (most are intra-speaker), but a change of speaker is often marked by a silence period.

2. A silence model was trained on the silence segments, and a 6-state ergodic HMM representing the global mean and variance of the speech was initialised on the speech segments.

3. A 2xERG6 HMM was created and initialised by coupling the ergodic HMM from step 2 with a 6-state A model from experiment 1. The 2xERG6 HMM was then trained on all the speech segments identified in step 1.

4. The fully trained 2xERG6 model was split into two 6-state ergodic speaker models. The A model was simply thrown away. The B model was used in conjunction with the SIL model from step 2 and a 24-state A model from experiment 1 to separate the speech data. The reason for using a 24-state speaker A model rather than a 6-state model as for speaker B was that the 24-state model was a more accurate speaker model.

Because there are no transitions between the two halves of a 2xERG6 model, each training segment only leads to the re-estimation of half the model parameters: the parameters corresponding to a single speaker. It is therefore important that step one succeeds in creating a segmentation where each segment only contains speech from one speaker.

*2.2.3 Experiment 3, Speaker Independent*

This experiment was basically a repetition of experiment 2, except that no pretrained models were used. Compared to experiment 2, this meant that in step 3 the 2xERG6 HMM was initialised directly from the speech data, and that in step 4 both sides of the 2xERG6 model were used as speaker models in the separation process.

### 2.3 Post Processing

Ergodic HMMs can not do duration modelling very well. This causes the recognition of a large number of very short "utterances", most of which do not mark a change of speaker, and many of which are misclassified (wrong speaker). To make up for this, the segmentations produced by Viterbi decoding the speech data, were post processed:

1. All consecutive utterances of the same type (i.e. speaker) were merged.

2. All SIL segments with a duration of less than 3.0 seconds (which were all SIL segments) were "removed" by extending the surrounding speaker segments to cover the SIL segment also.

3. Finally all sequences of segments shorter than 0.3 seconds were lumped into larger "compound" segments that were either at least $5 \cdot 0.3$ seconds long, or contained one sub-segment that was more than 0.3 seconds long. The compound segments were classified according to the type of sub-segment that made up the larger part of their duration.

After post processing the Viterbi segmentations, the new segmentations were effectively utterance level rather than word level segmentations.

## 3 EVALUATION MEASURES

Two measures of success were used. The first was the *sequence accuracy*:

$$\text{Seq. Acc.} = \frac{H - I}{N} \tag{1}$$

where H is the number of correctly recognised utterances (in terms of speaker id), I the number of insertion errors and N the total number of utterances. The second was the *duration accuracy*, which is the relative number of seconds of a recording that is correctly segmented:

$$\text{Dur. Acc.} = 1.0 - \frac{\text{total error duration}}{\text{total duration}} \tag{2}$$

## 4 RESULTS AND DISCUSSION

Table 1 summarises the overall results of experiment 1, 2 and 3 (including post processing).

| | Seq. Acc. (%) | Dur. Acc. (%) |
|---|---|---|
| **Exp 1, erg 1** | 84.9 | 92.9 |
| **Exp 1, erg 3** | 85.6 | 94.9 |
| **Exp 1, erg 6** | 89.3 | 95.9 |
| **Exp 1, erg 12** | 89.8 | 96.6 |
| **Exp 1, erg 24** | 85.9 | 97.3 |
| **Exp 2** | 88.8 | 95.8 |
| **Exp 3** | 86.6 | 93.5 |

Table 1: Overall sequence and duration accuracies obtained in the three experiments. In experiment 1, the number of states in the ergodic models was varied. In experiment 2, only one pretrained speaker model was used. In experiment 3, no information about the speakers was used in advance.

The baseline experiment showed that increasing the number of states in the speaker models increases the accuracy. The separation was, however, surprisingly good even when 1-state speaker models were used. The drop in sequence accuracy when going from 12 to 24 states in experiment 1 was due to an increased number of deletion errors at the post processing stage. Short segments were deleted, but overall the duration accuracy was increased.

Experiment 2 resulted in accuracies that were approximately the same as in experiment 1, when 6-state ergodic models were used. In experiment 3, the sequence and duration accuracies were respectively 86.6% and 93.5%, which is only slightly inferior to the corresponding baseline experiment (6-state ergodic models).

The duration accuracies were much higher than the sequence accuracies, because the conversations contained a large number of very short utterances that were difficult to detect ("Oh", "Yeah", "Huh", "Uh-huh", etc.), but which in terms of duration only made up a small fraction of the total conversations. Approximately 10% of the reference segments were less than 0.3 seconds long, and the duration accuracies can therefore not be expected to be much higher than 90%, since the post processing does not allow segments shorter than 0.3 seconds. These 10% of the reference segments made up less than 1% of the total duration.

The post processing step reduced the number of segments by well over 400%. An alternative to post processing would be to impose a penalty on transitions between models in the Viterbi decoding step. Probably this would achieve some of the same results, although it would still be necessary to process the (short) silence segments, which in most applications would be regarded as noise.

Figure 3 shows an alignment of the reference segmentation of the first half (150 seconds) of one of the conversations (full line) with the segmentation created in experiment 3 (dashed line). Silence periods have not been illustrated in either segmentation, and the exact misalignment of segment boundaries can therefore not be taken at face value (this has been compensated for in table 1).

# 5 CONCLUSIONS AND FURTHER WORK

The experiments showed that use of a 2xERG6 HMM is an effective way of creating speaker models when training data is not available (experiment 3). The 2xERG6 HMM is not sensitive to how it is initialised, but it is important that it is trained on speech segments that each contain speech from only one speaker. Such a training segmentation can be created by silence detection, and by requirering that all training segments be of short duration (here max 2.0 seconds).

An interesting question is how the results will generalise to conversations between more than two speakers. In principle it is straight forward to extend the approach by using a NxERG type HMM to create speaker models for N speakers. Presumably these models will be of poorer quality, but this might be compensated for by iteratively improving the models, for instance by using the first separation as the basis for training new speaker models.

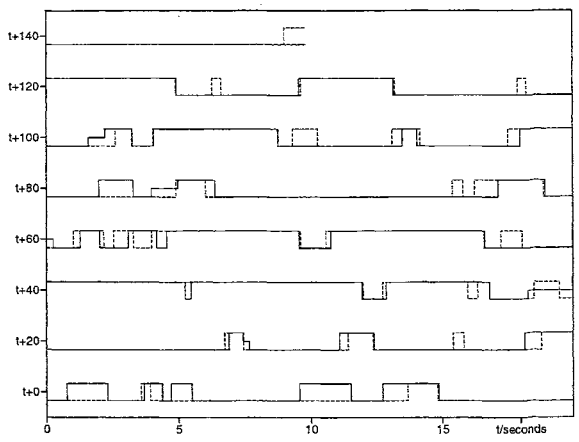


Figure 3: Experiment 3; Speaker Independent mode. "Binary" representation of respectively the reference segmentation (full line) and the segmentation created in experiment 3 (dashed line). "0" corresponds to speaker A, "1" to speaker B and "0.5" speaker A+B (only used in the reference segmentation).

# 6 ACKNOWLEDGMENTS

The experiments reported here were carried out as part of a M.Phil. thesis [6] at Cambridge University. I thank my supervisor then, Professor Steve Young, for his useful advice and critique during that project.

# References

[1] Debby Hindus and Chris Schmandt. Ubiquitous audio: Capturing spontaneous collaboration. In *CSCW 92 Proceedings*, pages 210–217, 1992.

[2] Steve J. Young. The HTK HMM toolkit: Design and philosophy. Technical Report CUED/F-INFENG/TR.152, Cambridge University, Engineering Department, 1993.

[3] Man-Hung Siu, George Yu, and Herbert Gish. An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In *IEEE ICASSP*, volume I, pages 189–192, 1992.

[4] Tomoko Matsui and Sadaoki Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. *IEEE Transactions on Speech and Audio Processing*, 2(3):456–459, July 1994.

[5] J. Goodfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *IEEE ICASSP*, volume I, pages 517–520, 1992.

[6] Jesper Ø. Olsen. Separation of speakers in audio data. M.Phil. Thesis, Cambridge University, Engineering Department, 1994.