Evaluation of a Channel Assignment Algorithm

Plutka T., Rothbucher M., Diepold K.

April 13, 2014

Abstract

In the course of the developement of an online teleconference system at the Institute for Data processing at the Technical University of Munich, a channel assignment algorithm for microphone arrays was presented [1]. This article evaluates the use of the channel assignment algorithm on usability in conference situations with dynamic speaker positions. For this purpose two experiments were created. The first one determines the average time until a change of the speaker position is fully processed by the algorithm. The second one is the processing of a four speaker conference with two conferees swapping positions.

1 Algorithm

The algorithm [1] to evaluate is used to do speaker channel assignment in telephone conferences. It combines SRP-PHAT and speaker recognition techniques to provide a more robust assignment of speech signals to the individual speaker channels. Recent evaluations focused on channel assignment in situations were participants did not move. A major point of this article is to evaluate the channel assignment algorithm in conference situations with speakers changing places.

1.1 Algorithm Revisions

During the evaluation, several changes to the original algorithm were done and combined in two different revisions. The first revision holds several optimizations concerning computing time in offline processing, whereas the second revision introduces a buffer to the model adaption process. This buffer improves the ability to recognize a speaker that has changed its position, for example from the table to a blackboard.

1.2 Algorithm Overview

As the optimizations to improve computing time don't affect the original sequence of the algorithm, the overview is basically the same for all existing versions. First, section 1.3 will give a short description of the whole algorithm and more detailed insight into every stage afterwards. The Buffer, which was introduced in revision 2 is explained in section 1.5.

1.3 Original Algorithm

To provide a better understanding, the algorithm can be divided into 8 sections, which are of different complexity. An Initialization stage prepares models, matrices and audio files for further processing. The SRP-PHAT localizer and GSS Module can be run in parallel in a real-time implementation and provide the input to the geometry stage, which compares localization data against trained models. For higher reliability, speaker features of the extracted streams are checked against the model database. If results differ, speaker recognition beats the localization stage. A DER calculation is only used for evaluation and not part of a possible real-time implementation, as the ground truth will not be known. The last stages are a verification step, that adapts speaker models and the output stage, which prepares output streams for every speaker.

initialization

Figure 2 visualizes the initialization process. After loading the default variables, the existance of the speaker model files is checked. If no model files are found new ones will be trained from audio files with a single speaker. In the next step, ground truth and audio files are read and the ground truth is copied to a new table. The audio files are windowed using a hamming window function.

SRP-PHAT + GSS

A Steered Response Phase Trasform (SRP-PHAT) Algorithm is used to locate sound source positions, afterwards these are processed by Geometric Source Separation (GSS). This block is fed with the enframed eight channel audio streams from the microphone array. Data from GSS output will be assigned to different speaker streams by geometry, feature and decision stage.

geometry stage

The geometry stage shown in 3 first calculates the origin of a speech utterance in spherical coordinates. Afterwards, source position of the speech utterances is compared to positions saved in the speaker models. Using a winner takes it all approach, the model with minimal distance to the speech utterance is chosen as active speaker. It is absolutely important to distinguish between localization by the SRP-PHAT algorithm and localization

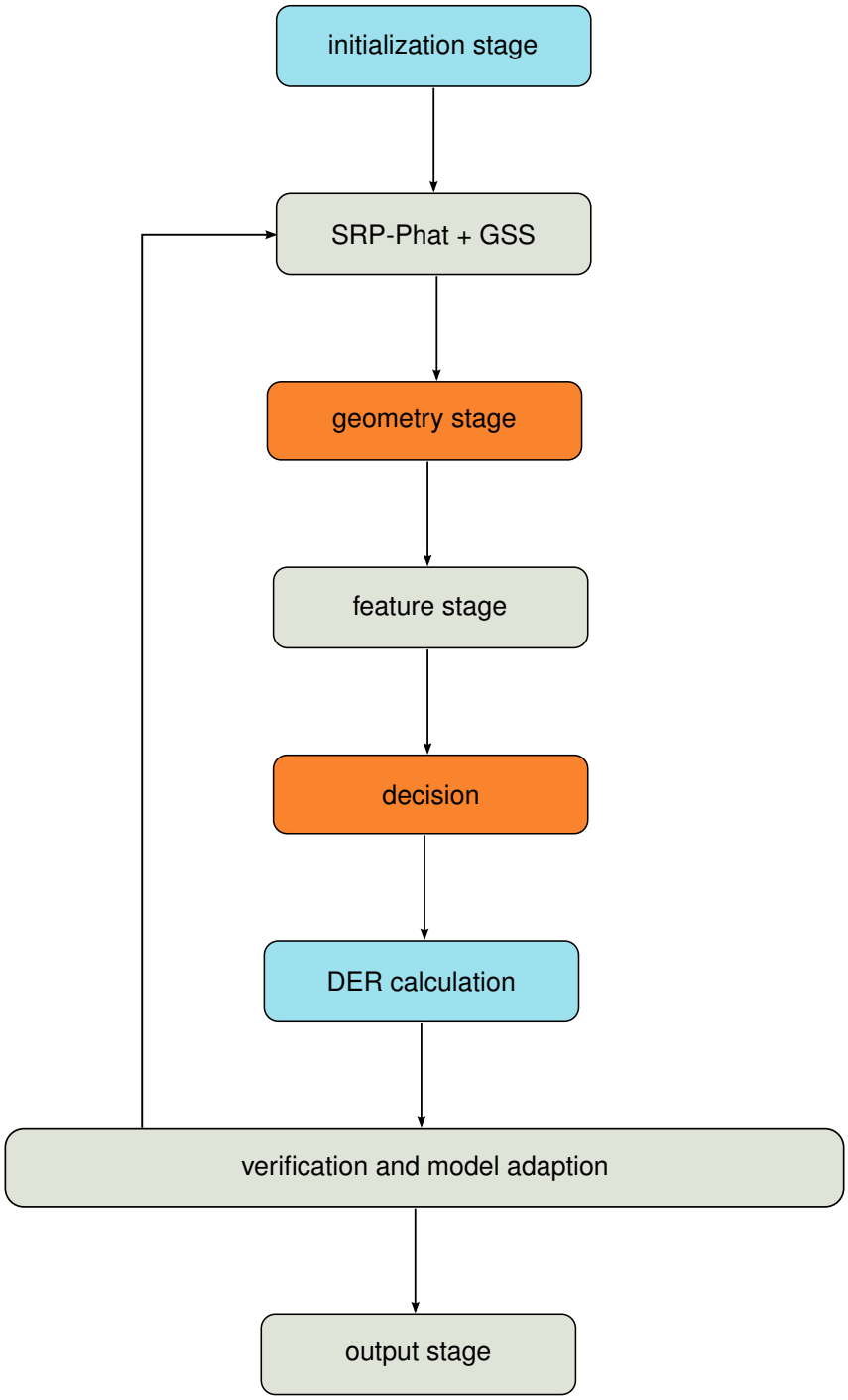

Figure 1: Overview

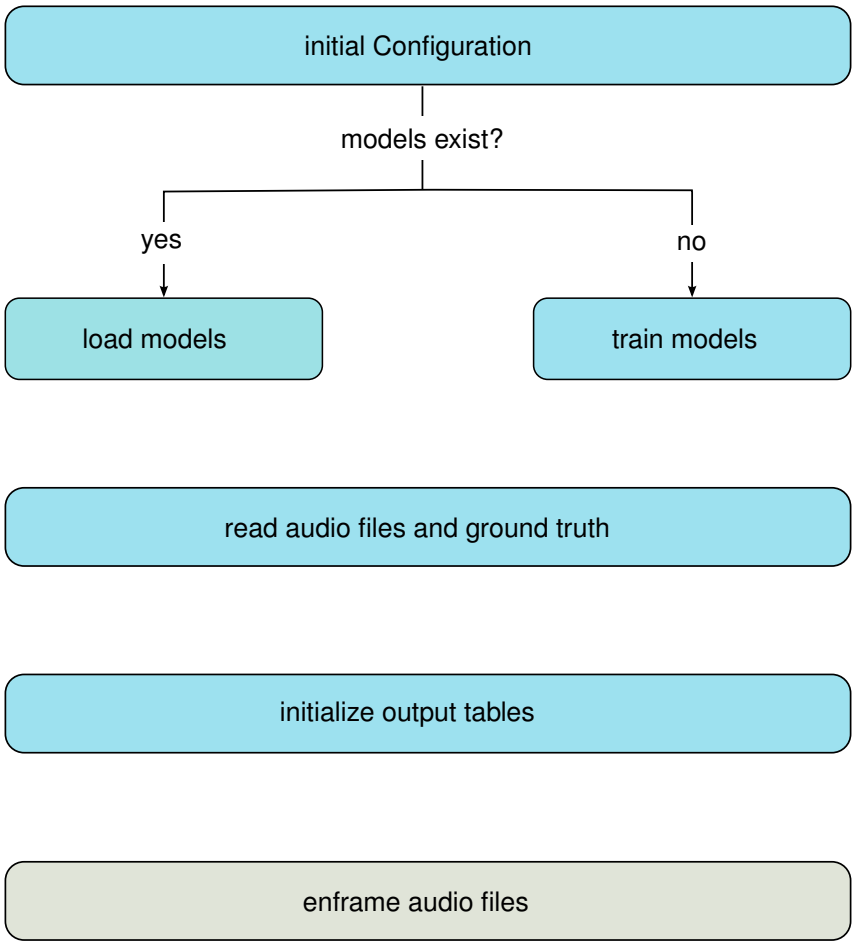

Figure 2: initialization stage

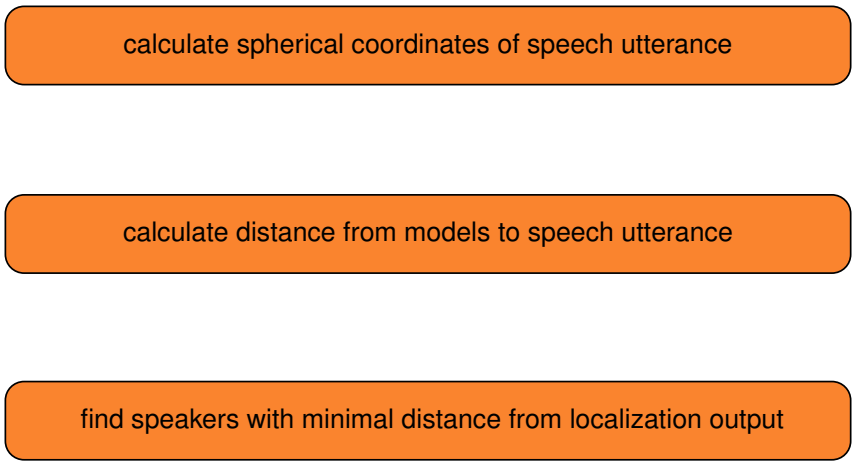

Figure 3: geometry stage

of the channel assignment. Output of the geometry stage and hence that of the algorithm is the position saved in the speaker models.

feature stage

Feature stage (Figure 4) is used to support, or correct the results of the geometry stage. Speaker features of every stream are calculated and compared against these of the speaker models using a maximum likelihood algorithm.

decision stage

The flow diagram of the decision stage is shown in figure 5. If the model position of the localized speaker model deviates more than 10° from the localized speech utterance, speaker recognition is used to assign the source to the speaker channel with maximum likelihood calculated by the feature stage.

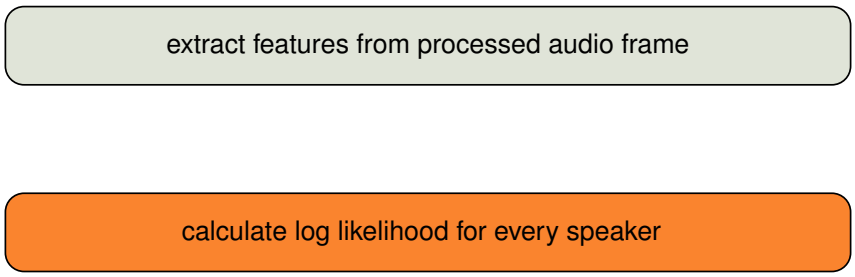

Figure 4: feature stage

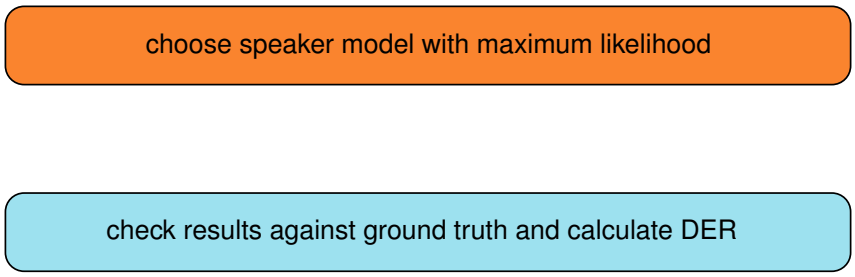

Figure 5: decision stage

verification stage

The Verification stage ensures, that the speaker models are constantly updated and even makes it possible to keep track of speakers changing positions. If only one speaker is active, and model position and localization data by the SRP-PHAT algorithm do not deviate more than 15°, the MFCC's of the active speaker model are updated. A buffer stores the speech signal, if the active speaker is the same as in the last time step. If one second of speech is collected, the algorithm checks if model position and mean of the SRP-PHAT localization are consistent. If not, a counter is increased. After reaching a threshold, the model position is adapted to the SRP-PHAT localization position. Output file preparation assigns the samples to the corresponding speaker streams. A detailed illustration of the process is shown in figure 6.

1.4 Improved Original Algorithm

Changes done to the original algorithm were mostly addressing offline processing time in MATLAB, as computing of conferences of about 10 minutes took over 9 hours. These changes are not relevant for an evaluation of the algorithms performance in channel assignment, but mentioned here for the sake of completeness.

1.4.1 Changes

1. Inserted extra columns to the xls generation, to document, which speaker was localized and which speaker was actually recognized by the speaker recognition. This will be used later to illustrate the process of adapting the model position when speakers change their position.

2. Disabled that the adapted speaker model is saved after every adaption, as the matlab internal 'save' command is very time consuming. Models are now written to a new file after the audio streams were fully processed.

3. Disabled that the output of the speaker localization is written to a '.mat' file after

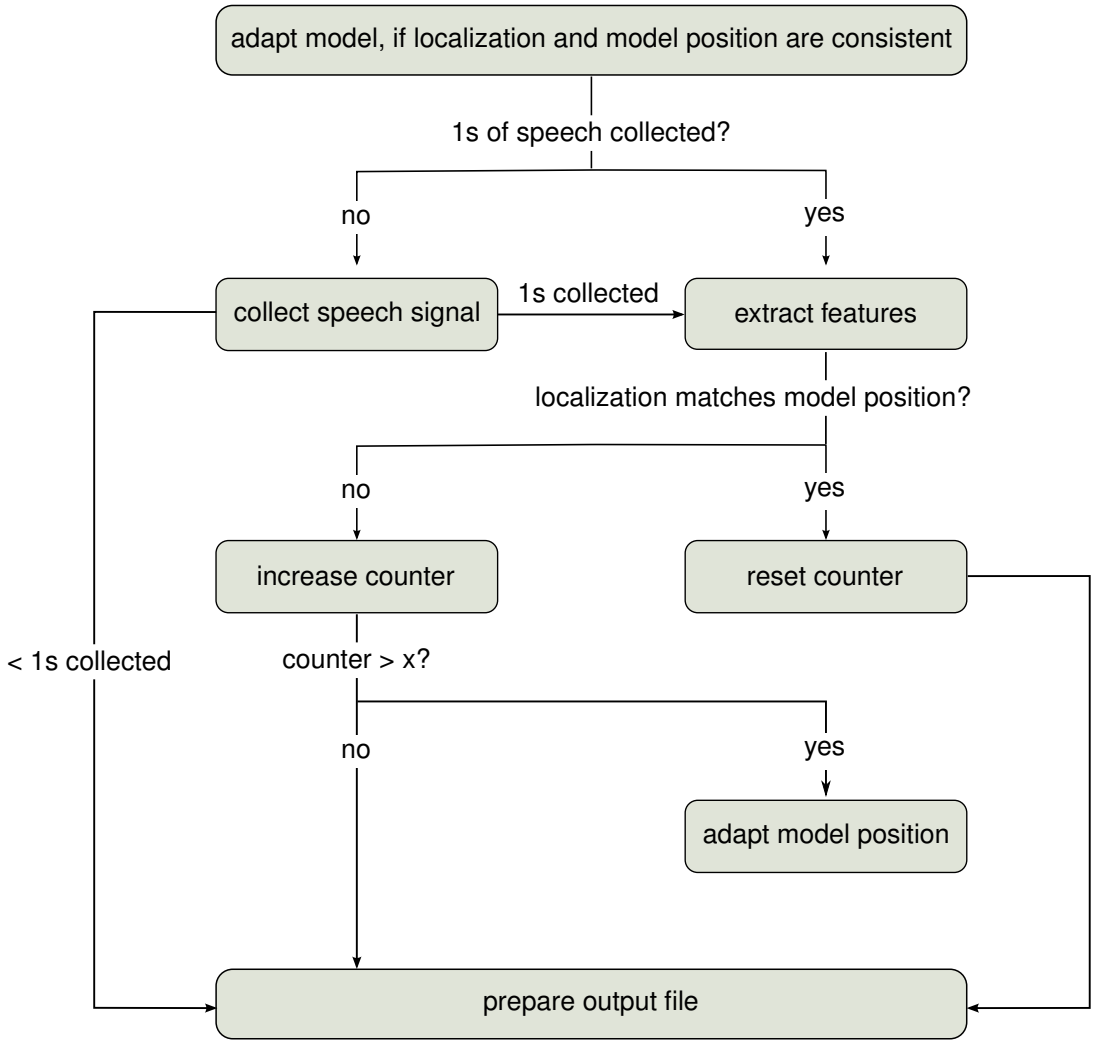

Figure 6: verification stage

Videolab dimensions	6.3m x 4m x 2.8m				
frequency bin in Hz	250	500	1000	1995	3981
Videolab reverberation time t_{60} in s	0.2545	0.2169	0.2230	0.2466	0.2149

Table 1: Room characteristics of the videolab

every frame, as it contains only one frame and the old output file is overwritten every time.

1.5 Buffered Version

A major problem that occurred with the original implementation is, that false detections in single frames reset the counter, which is responsible for the model position adaption. If a speaker changes position during a conference, false detections delay the correction of the speaker model position and therefore raise the number of samples assigned to the wrong speaker channel. To challenge this problem, a buffer was introduced. This buffer stores the speech signal collected until frame [n], if frame [n-1] was assigned to another speaker. If the chosen speaker in frame [n+1] is same as in frame [n-1], the buffer is written back to be used by the evaluation stage.

2 Evaluation

Evaluation is divided into two parts. As the processing of conferences with speakers on fixed positions was already evaluated by [1] we will focus on scenarios with speakers changing positions during the conference. In the first part, a single speaker scenario is analyzed. The second part is a four speaker conference with two participants swapping places.

2.1 Audio recordings

All conference files were recorded at the videolab of the Institue for Data Processing. Table 1 shows the room characteristics of the videolab. Sampling frequency was at 48kHz. The speaker recordings used to simulate the conferences were made in [Arbeit von Korbi], the conference files itself were created especially for this evaluation.

2.2 Single speaker

In order to test the algorithms ability to correctly assign speakers who changed places during a conference, a special single speaker scenario was created. The algorithm was trained on three speaker models who were placed in the room, but only one of them was

Buffer	Threshold	DER	Average time until model is adapted
yes	3	2.01	2.72s
no	3	2.22	7.54s
no	7	4.84	17.84s

Table 2: Single speaker experiment results averaged over 41 trials.

actually talking. After 60 seconds, the conferee changed its position. Goal of this experiment was to measure how long it took the algorithm to correct the position of the speaker model. In total 41 recordings with 11 different speakers were analyzed with both revisions of the algorithm. The unbuffered algorithm used two different thresholds, the buffered Version only one. Average times until the speaker position in the models was corrected are given in table 2. It can be seen, that the buffered version is about 3 times faster in recognizing the speaker changing its seat. Difference in DER with or without the buffer is only about 0.2%. This is because model position and localization by the SRP-PHAT stage differ more than 10° after the speaker has changed place and therefore localization is overridden by the speaker recognition until the model position is corrected.

2.3 Videolab Conference

After having tested the isolated case of one speaker changing seats, a four speaker conference was created to show the algorithms abilities in a more general scenario. Three male an one female speaker were placed around a microphone array in 1.3m radius and 45° distance between speakers. Length of the conference was 7:53 minutes. If speakers were active, speech signals were between 4s and 28s long. The conference was recorded in two different variations. In the first one, speakers kept their seats, whereas in the second one, two speakers (Jonas and Kathrin) swapped their seats after 256s. Table 3 contains information whether model correction was turned on or off and the threshold values until the corresponding speaker model was adapted. The last column shows how long it took the algorithm to correct the first model position. Time until the second model is corrected is not necessarily relevant, as the position where it is after the change will be empty when the first model is corrected. So, the localization is not able to assign the speech utterance to a model, but the speaker recognition assigns the utterance correctly. After reaching the threshold the second model is corrected too. Figure 7 shows the Videolab Conference. The vertical line denotes the point where the change happened. Right after the speakers swapping seats, the grey stream contains data that should be in the red one. But after about 3s the red model is adapted and the audio signal is now assigned correctly. As mentioned before, there is no model at the position where the grey speaker is now. So localization is outruled by the speaker recognition, and the grey stream is assigned properly.

If all conferees remained on their position, or model correction was turned on, DER was

Speakers change place	Model correction	Threshold	DER	Time until first model is adapted
no	yes	2	5.182	-
no	yes	3	5.125	-
yes	yes	2	5.190	3.402s
yes	yes	3	5.770	5.066s
no	no	-	5.108	-
yes	no	-	27.0525	-

Table 3: Videolab conference results.

about 5%. But if model correction is turned off, DER increased to about 27%, as soon as speakers swapped places. When model correction is turned on, and threshold is set to 2, there is practically no difference in DER between conferences where speakers keep their positions the whole time. Reducing the threshold, time until the first model is adapted goes down to 3.402s.

3 Conclusion

The experiments with a single speaker and the conference scenario confirm, that a use of the channel assignment algorithm in conferences with participants who change positions is basically possible and will provide good results. Combination of speaker recognition and SRP-PHAT significantly lowers the DER, compared to using only one of both, but still can be improved. Introducing a reliability scale for speaker recognition and localization that dynamically determines which of the two choses the active speaker, could be a reasonable addition to fasten and stabilize the model adaption process.

References

[1] K. Steierer. Teleconference channel assignment. Diploma thesis at the Institute for Data Processing, Technische Universität München, June 2013.

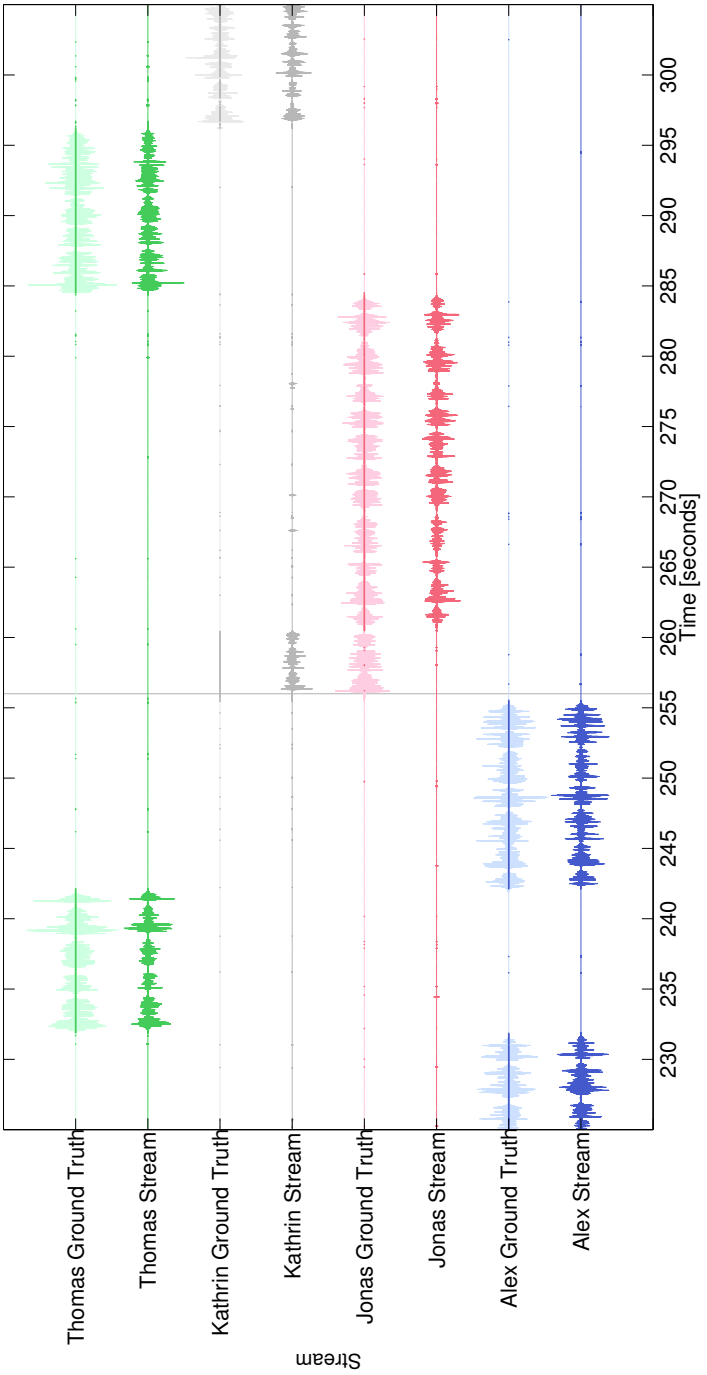

Figure 7: Output streams of the Videolab Conference. Light colors denote the ground truth. Corresponding darker colors the speech data assigned to the audio stream.