Automatic Annotation of Conversation in Large Multimodal Data Sets by Enhanced Voice Activity Detection and Throat Microphones

B. Tessendorf[1], P. Derleth[2], M. Feilner[2], D. Roggen[1], T. Stiefmeier[1], G. Tröster[1]

[1] *Wearable Computing Lab., 8092 Zürich, Schweiz, Email: {lastname}@ife.ee.ethz.ch*

[2] *Phonak AG, 8712 Stäfa, Schweiz, Email: {firstname.lastname}@phonak.ch*

Introduction

The accurate annotation of multimodal data sets with the application-specific ground truth is the foundation for statistical analyses and for applying supervised machine learning algorithms. These techniques rely on annotated reference data sets[1] to build a model of the training data and to quantitatively evaluate the resulting classification performance on the test data. However, obtaining annotations is challenging, especially for large and complex data sets. Manual annotation is a cumbersome, error prone, and extremely time consuming process [1]. There is a need for reliable methods for automatic annotation. We use throat microphones for annotation as they are well suited to detect own voice and thereby avoid challenges of environmental noise, overlapping speech and speaker diarization. We investigate how to process voice activity data recorded with a throat microphone on multiple participants to automatically annotate conversation. The contributions of this paper are: (1) demonstrating that we can accurately recognize entire phases of conversations by enhancing VAD with a conversation filter, (2) characterizing speaker-dependency, and (3) a comparison of our approach to basic VAD. A conversation is characterized by a turn taking interaction between two or more conversation partners. As such, detecting conversation is more complex than detecting the occurrence of speech only, which is traditionally achieved with VAD. We describe how to enhance VAD with a conversation filter to automatically annotate conversations. This speeds up the annotation process of acquiring reference datasets in comparison to the current manual annotation. We assume that the conversation partner is actively involved in the conversation.

Automatic Annotation of Conversations

Throat microphones pick up vibration on the user's neck. This makes them inherently robust to any acoustical noise and well suited for tasks that require reliable annotations such as the acquisition of reference data sets. They can be used to annotate synchronized data of the remaining modalities in the reference data set. The throat microphone needs to be worn by the conversation partner of interest only during the acquisition of the reference data set. We compare three approaches to automatically annotate conversation based on the recorded data. First, we consider a basic VAD as a baseline approach. We enhance the baseline approach with a consecutive median filtering step, referred to as conversation filter 1 (CF1). The third approach extends CF1 by introducing speaker-dependent threshold values, referred to as conversation filter 2 (CF2). Fig. 1 illustrates the processing steps.

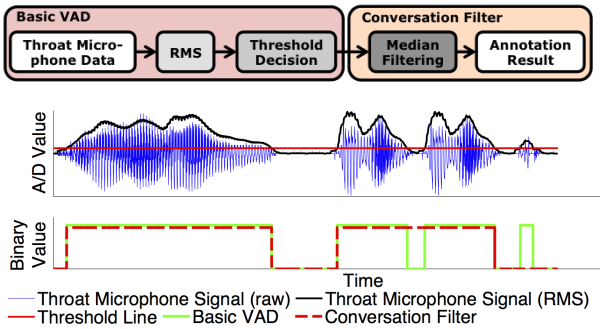

Figure 1: Block diagram and example data to illustrate the automatic annotation of conversation. VAD is performed on the raw throat microphone data. The conversation filter merges contiguous parts and removes insertions.

Algorithms We implemented basic VAD based on the time energy of the signal [2]. As a feature we calculate the root mean square (RMS) on a sliding window with a fixed step size of 1 second on the raw throat microphone data. Based on a threshold analysis of the time energy of the signal we then obtain an indicator of voice activity. The optimal threshold value is found by applying a decision stump [3] and a 10-fold cross validation. We implement the speaker-independent conversation filter CF1 and speaker-dependent conversation filter CF2 by median filtering the VAD signal to obtain the final annotation result. Median filtering preserves edges and suppresses noise. This way we merge speaking pauses during conversations and events that are too short to constitute a conversation situation. We consider fixed system parameters for all participants (CF1) and speaker-dependent values (CF2).

Evaluation Data Set We evaluate our approach for automatic annotation on the multimodal data set described in [4]. Throat microphone data was recorded with 8 bit at 8 kHz using an Alan AE38. The data set is balanced and comprises two classes: For each point in time it was manually annotated if the participant is having a conversation by inspecting video footage. The given data set has a total length of over 6 hours and comprises real conversation situations for 3 females and 5 males in the age of 24–59. For this explorative data set manual annotation

[1] A reference data set is used in the development phase of signal processing or machine learning methods to identify algorithms and their parameters. Thus, it is possible to instrument subjects with a larger number of sensors than is the case during the actual operation of the system. During operation only a minimal set of selected sensors is used, with the algorithms and parameters identified from the reference data set.

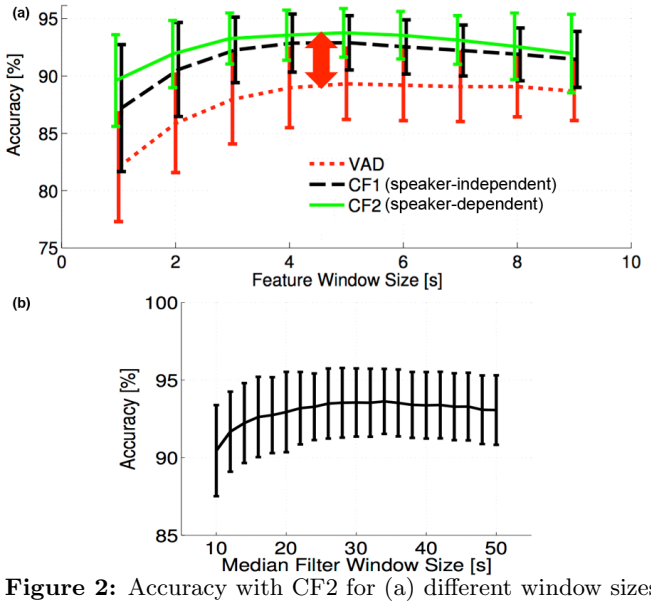

Figure 2: Accuracy with CF2 for (a) different window sizes and optimal median filter window size and (b) different median filter window sizes and a feature window size of 5 seconds. Values are averaged over all participants and the standard deviation is given.

was possible. However, for large-scale recordings this is not feasible in a reasonable amount of time. We evaluate the automatic annotation on this data set to characterize its performance. We can then deploy the automatic annotation method in future larger-scale reference data sets without using manual annotation.

Evaluation Method We compare our annotation results with the ground truth annotation of the data set that was obtained by manual annotation. For evaluation we use continuous evaluation measures introduced in [5]. They are well-suited for our evaluation task as they allow a fine-grained categorization of classification errors into severe and less severe errors by capturing common artifacts found in continuous activity recognition, including insertions (INS), merges (MER), deletions (DEL), fragmentations (FRA), underfills (UND), and overfills (OVE). We calculate the accuracy (ACC) as the share of samples where the annotation result matches the ground truth. We assess the influence of the RMS feature window size and the median filter window size on the annotation accuracy.

Results and Discussion

Typical conversation situations comprise periods with and without voice activity such as pauses, or clearing throat as depicted in Fig. 1. A basic VAD approach is not sufficient to model this structure. Table 1 shows the resulting continuous evaluation measures for VAD and the conversation filters CF1 and CF2 with a feature window size of 5 seconds. An average accuracy of 89.3% was achieved with VAD. It is improved to 92.9% by using CF1. The main causes for the lower performance of VAD are fragmentations and insertions, together 7.1%. With CF1 the influence of these error sources drops to 0.6%. The improvement results from implicitly admitting only plausible durations of conversation activity. The conversation filter merges speaking pauses during conversations. Events that are too short to constitute a conversation situation like clearing throat or coughing are removed. The most important remaining error sources are underfill and overfill, together 6.4%. Underfill and overfill errors are not severe when annotation jitter can be tolerated. Merge and deletion errors are insignificant for all variants. Fig. 2 depicts the accuracies for VAD and the conversation filters CF1 and CF2 for values of the feature window size in the range of 1 to 9 seconds. The arrow indicates the gain in accuracy achieved with conversation filters. With optimal parameters for each participant in CF2 the average accuracy improves to 93.9%. Reasons for the observed user-dependency are the placement of the throat microphone at the neck, the strength of the voice and being more talkative or passive in a conversation. Overall CF2 shows best performance with a feature window size of 5 seconds and a median filter window size of 34 seconds. Fig. 2 (b) depicts the accuracies in function of the median filter window size for CF2 and a feature window size of 5 seconds.

	ACC	PREC	REC	FRA	MER	UND	OVE
VAD	89.3 (3.3)	85.6 (5.9)	90.7 (6.2)	3.1 (2.3)	4.0 (2.0)	0.8 (0.5)	2.7 (1.2)
CF1	92.9 (2.5)	89.7 (5.3)	94.7 (4.4)	0.2 (0.4)	0.4 (1.1)	2.0 (1.6)	4.4 (2.2)
CF2	93.9 (2.2)	90.8 (4.4)	95.4 (3.1)	0.4 (0.6)	0.4 (1.2)	1.5 (1.1)	3.7 (0.6)

Table 1: Continuous evaluation measures for VAD and the conversation filters CF1 and CF2 given in %. INS and DEL are near zero and not shown here. Values are averaged over all participants and the standard deviation is given in brackets.

Conclusion

A basic VAD approach is not sufficient to accurately annotate entire phases of conversations because it does not consider speaking pauses and artifacts. However, by enhancing VAD with a consecutive conversation filter the average annotation accuracy is increased from 89.3% to 92.9%. With speaker-dependent parameters the average accuracy is further optimized to 93.9%. The system is robust to variations of the feature and filter window sizes, and also concerning speaker dependency and environmental noise. It represents a computationally lightweight and low-cost solution deployable in mobile settings.

References

[1] D. Roggen et. al., "Collecting complex activity data sets in highly rich networked sensor environments," in *7th Int. Conf. on Networked Sensing Systems*, 2010.

[2] G. Carli and R. Gretter, "A start-end point detection algorithm for a real-time acoustic front-end based on dsp32c vme board," in *ICSPAT*, 1992.

[3] W. Iba et. al., "Induction of one-level decision trees," in *Conference on Machine Learning*, 1992.

[4] B. Tessendorf et. al., "Recognition of hearing needs from body and eye movements to improve hearing instruments," in *Conf. on Pervasive Computing*, 2011.

[5] J. Ward, P. Lukowicz, and H. Gellersen, "Performance metrics for activity recognition," *Transactions on Information Systems and Technology*, 2011.