Automatic Voice Activity Detection in the Multilingual UTEP-ICT Cross-Cultural Multi-party Multi-modal Dialog Corpus

Katherine Sittig-Boyd
Simmons College
Boston, MA 02215
sittig@simmons.edu

Elaine Short, Maja Matarić
University of Southern California
Los Angeles, CA 90007
{elaine.g.short,mataric}@usc.edu

***Abstract*— We present a novel approach to speech identification in group interactions with individual microphones, extending Moattar and Homayounpour's algorithm (2009) to identify individual speech even when background speakers are present. Using this approach, we annotated speech across the three language groups of the UTEP-ICT Corpus. Comparisons with human-coded ground truth using Cohen's Kappa for inter-rater reliability resulted in a mean score of .77 (s=.23). This method enables individual speech detection in multi-party interactions without directional microphones.**

I. Introduction

Identifying the primary speaker in a multiparty interaction is a crucial aspect of enabling autonomous agents, such as virtual agents or social robots, to interact in multiparty scenarios. Using voice activity detection (VAD), we introduce an approach to discriminate which participant is speaking, without necessitating strongly directional microphones.

Due to the nature of multiparty interactions, there are often multiple participants speaking at one time. However, some instances of speech, such as backchannel (verbal interjections such as "yes" or "uh-huh" from another participant in the interaction which are meant to show agreement or understanding but do not contribute to the conversation), may be unimportant with respect to identifying who the primary speaker is in a multiparty interaction.

We introduce an altered version of Moattar and Homayounpour's VAD algorithm [6] to account for voice activity not originating from a microphone wearer , during the course of a multiparty interaction. To test the accuracy of our speaker detection method, we used this VAD method to identify and annotate instances of speech in the audio recordings from the UTEP-ICT Corpus [4].

In the next section, we discuss similar work, as well as specific outcome goals. In Section III, we discuss the corpus data, particularly with regard to audio quality, and describe the annotation algorithm. In Section IV, we discuss the reliability of our method in comparison to human-coded annotations. Section V outlines future work related to this project, including ongoing dialogue pattern analysis, as well as applications for this work in multiparty interactive scenarios.

II. Background and Motivation

A. Voice Activity Detection

One critical aspect of speech and audio processing is voice activity detection (VAD), which identifies audio features specific to human-produced vocalizations. Several approaches to VAD exist, including deep belief networks [9], long-term signal variability [3], [10], and spectral clustering [7]. However, the UTEP-ICT Corpus audio contains relatively little background noise, so a feature selection algorithm such as [6] is sufficient for our purposes.

Fig. 1. American 2 group, Naming Task.

B. UTEP-ICT Corpus Multiparty Interactions

The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus [4] captures the interactions of twelve groups with participants from three different native language backgrounds, Arabic, American English, and Mexican Spanish, with the goal of identifying communicative patterns with respect to culture in each interaction. Groups of four participants from the same language background are given five interactive tasks, each lasting approximately 10 minutes, resulting in a total of 50 minutes recorded for each participant group. These tasks were as follows:

1) Describe your pet peeves
2) Figure out a movie you have all seen; discuss its best and worst parts
3) Determine a good name for a toy
4) Tell a story about the same toy
5) Describe an inter-cultural experience

While proxemics, gaze, and turn-taking have been coded for each of the interactions, speech remains largely unannotated, with the exception of three of the American tasks, across the four participants. The multilingual nature of the corpus increases the difficulty of the annotation task, since coders may not be capable of understanding Mexican Spanish or Arabic enough to identify speech originating from the primary speaker, especially if there is voice similarity between two speakers.

Corpora comprised of multiple speakers are often gathered using directional or lapel microphones [8], [5], [1]. Although the audio quality may be good, there is often background speech from other participants. In the UTEP-ICT Corpus, although each participant is fitted with a lapel microphone, the recordings for individual speakers still contain voice activity originating from other participants. Although human coders have the capacity to distinguish among multiple speakers in a single recording, VAD algorithms are not as discriminatory.

Additionally, the multiparty nature of these interactions resulted in over 30 hours of audio data due to the individually recorded participants, which is time-consuming to annotate by hand. A VAD-based annotation method which is capable of identifying speech instances across multiple languages provides a means of analyzing speech patterns related to turn taking, percentage of time speaking, and balance of speakers in group interactions.

III. Methods

A. Data

For each group, the participants' speech was recorded using individual wireless, pin-on lapel microphones. Speakers were labeled as "A", "B," "C," and "D." The full audio recordings were split according to the five tasks. During two tasks, the recording equipment for one participant malfunctioned, and the audio was not reconstructed from other recordings. Without these recordings, there remained 178 audio files across all groups, tasks, and participants.

B. Algorithm

We provide an overview of Moattar and Homayounpour's algorithm as follows: Given 10 ms frames of audio samples,

1) Calculate the initial energy of the frame;
2) Identify the fundamental frequency of the frame;
3) Obtain the spectral flatness measure (SFM) of the frame.

These measurements are then compared to energy, fundamental frequency, and SFM thresholds as outlined in the algorithm. If at least two features cross these thresholds, the frame is labeled as a "sound" frame. Otherwise, it is labeled as "silence." Five or more consecutive "sound" frames indicate voice activity, while ten or more consecutive "silence" frames designate a lack of voice activity.

To counteract this algorithm's tendency to pick up on any instances of speech, originating from any speaker, in a multiparty interaction, we introduce a root mean square calculation to measure the signal power contained in an audio frame. We calculated the mean RMS measure in each 10 ms audio sample in the human-coded speech annotations to determine an RMS threshold, as seen in 1.

$$MinimumRMS = 400. \tag{1}$$

In our modified algorithm, an audio sample is only tested for speech-like features if it crosses the RMS threshold. This allows for the detection of the primary speaker even if another participant is talking in the background.

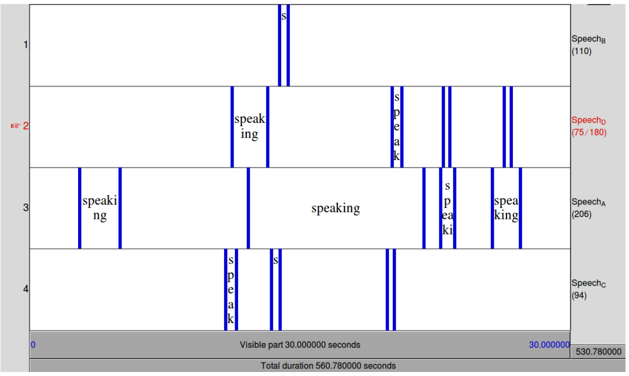

Fig. 2. Annotations of speakers A, B, C, and D in four Praat TextGrid tiers. The shorter annotations may indicate backchannels, such as ”mm-hm” or ”yes” from listening participants.

IV. Results

Using the algorithm described prior, we generated automatic annotations on the audio files which had already been human-coded, resulting in 22 automatically annotated files for comparison against the human-coded ground truth annotations.

A. Inter-rater Reliability

Calculating Cohen's Kappa [2] to compare the computer-coded annotations with the hand-coded annotations resulted in a mean Kappa score of .77 ($SD = .23$). The audio recordings of the American 3 tasks are comprised of audio re-created from the other participants' recordings to correct for microphone failure mid-task, which may account for the slightly lower Kappa scores.

V. Discussion

A. Contributions

1) Annotations: After ensuring the reliability of the automatic annotation method, we applied our algorithm to 178 audio clips across all 12 groups, 48 speakers, and 5 tasks, producing annotations in both ELAN and Praat. These annotations will be used to analyze conversational patterns among the cultural groups, including considerations for patterns in turn taking, percentage of time spent speaking, and speaker balance among the four speaking participants.

2) Voice Activity Detection for Multiparty Interactions: In multiparty interactions between a group of human participants and an autonomous agent, such as a robot or virtual agent, this method can be used to identify when one of the participants is speaking.

TABLE I
KAPPA SCORES

American 1 Naming Task	
Speaker A	.79
Speaker B	.85
Speaker C	.84
Speaker D	.91
American 1 Story Task	
Speaker A	.88
Speaker B	.95
Speaker C	.94
Speaker D	.93
American 2 Naming Task	
Speaker A	.85
Speaker B	.86
Speaker C	.78
Speaker D	.92
American 2 Story Task	
Speaker A	*
Speaker B	.80
Speaker C	.82
Speaker D	.92
American 3 Naming Task	
Speaker A	*
Speaker B	.62
Speaker C	.48
Speaker D	.16
American 3 Story Task	
Speaker A	.73
Speaker B	.68
Speaker C	.76
Speaker D	.46

B. *Limitations*

This algorithm does not discriminate laughter from speech; using this VAD approach in multiparty interactions which necessitate identifying participants’ sentences and questions may not be applicable. Additionally, despite the intensity filtering, loud background instances of noise (eg, group laughter, a participant shouting) may get incorrectly labeled as speech produced by microphone wearer. In the course of annotating the “naming” and “story” tasks, the toy’s song was occasionally mislabeled as “speech,” since it was initially recorded by a person, and produced noise that not only passed the VAD thresholds but also reached the necessary noise intensity level to indicate an occurrence of human speech.

VI. CONCLUSIONS

This method is applicable in multiparty communicative scenarios in which all participants are individually fitted with microphones, and does not necessitate microphones which are strongly directed. Annotations generated automatically by our method of voice activity detection had an inter-rater agreement with the hand-coded annotations of .77, using Cohen’s Kappa. The annotations created using our method will be used to analyze dialogue patterns cross-culturally in the UTEP-ICT Corpus.

Going forward, we aim to include a feature that will distinguish laughter from speech, since these two modes of verbalization serve different communicative purposes.

ACKNOWLEDGMENTS

The author would like to thank Dr. Maja Matarić, Elaine Short, the members of the Interaction Lab, CRA-W, and the organizers of the DREU program. This work is supported by the National Science Foundation (Expeditions in Computing, CNS-0709296; Human-Centered Computing (HCC) Program, IIS-0803565).

REFERENCES

[1] Aran, Oya, Hayley Hung, and Daniel Gatica-Perez. ”A multimodal corpus for studying dominance in small group conversations.” Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010 (2010): 22.

[2] Cohen, Jacob. ”A coefficient of agreement for nominal scales.” Educational and psychological measurement 20.1 (1960): 37-46.

[3] Ghosh, Prasanta Kumar, Andreas Tsiartas, and Shrikanth Narayanan. ”Robust voice activity detection using long-term signal variability.” Audio, Speech, and Language Processing, IEEE Transactions on 19.3 (2011): 600-613.

[4] Herrera, David, et al. ”The UTEP-ICT cross-cultural multiparty multimodal dialog corpus.” The Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. 2010.

[5] McCowan, Iain, et al. ”The AMI meeting corpus.” Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research. Vol. 88. 2005.

[6] Moattar, M. H., and M. M. Homayounpour. ”A simple but efficient real-time voice activity detection algorithm.” Signal Processing Conference, 2009 17th European. IEEE, 2009.

[7] Mousazadeh, Saman, and Israel Cohen. ”Voice activity detection in presence of transient noise using spectral clustering.” Audio, Speech, and Language Processing, IEEE Transactions on 21.6 (2013): 1261-1271.

[8] Oertel, Catharine, et al. ”D64: A corpus of richly recorded conversational interaction.” Journal on Multimodal User Interfaces 7.1-2 (2013): 19-28.

[9] Zhang, Xiao-Lei, and Ji Wu. ”Deep belief networks based voice activity detection.” Audio, Speech, and Language Processing, IEEE Transactions on 21.4 (2013): 697-710.

[10] Tsiartas, Andreas, et al. ”Multi-band long-term signal variability features for robust voice activity detection.” INTERSPEECH. 2013.