Audiovisual Voice Activity Detection and Localization of Simultaneous Speech Sources

Vicente Peruffo Minotto[1],
Advisor: **Dr. Claudio Rosito Jung**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500 – 91501-970 – Porto Alegre – RS – Brazil

{vpminotto,crjung}@inf.ufrgs.br

1. Introduction

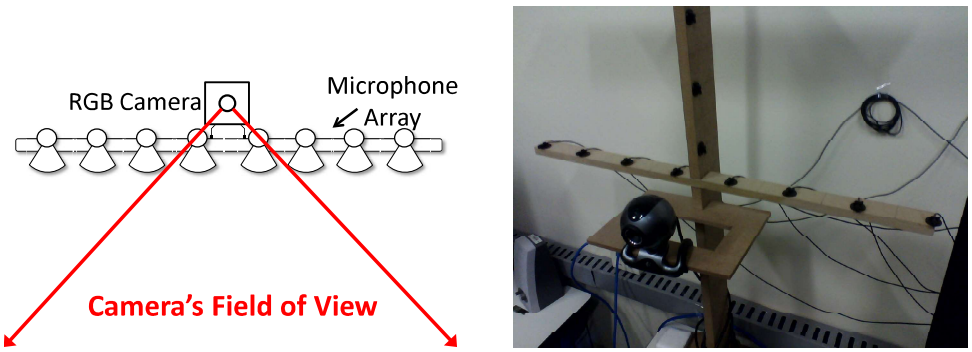
When a computer is used for a generic task, a mouse and a keyboard are predominantly employed as interfaces between the user and the machine. Despite being popular, they may not be adequate for a variety of applications, implying that other ways of human-machine interaction might be more promising (such an example is the touchscreen in tablets and smartphones). Combining this fact to the constant advances of computer technologies, it is natural to exist high interests in developing digital forms of interaction that are similar to those of common use between humans. Particularly, speech represents a vast part of the information exchanged during those interactions [Jaimes and Sebe 2007]. Therefore, once computers are able to efficiently comprehend human-like communication, human-computer interaction (HCI) becomes more convenient and effective.

However, differently from human-human interaction, HCI still presents many challenges. As an example, in automatic speech recognition (ASR), which is one of the main branches of HCI [Thiran et al. 2010], it is necessary to recognize words in audio signals that may have been corrupted, such as by environmental noise, reverberation and other competing speech sources. Therefore, to compensate for these degradations, user-level systems require front-end techniques to function robustly. To diminish this problem, Voice Activity Detection (VAD) and Sound Source Localization (SSL) arise as two of the most important preprocessing tools in speech-based HCI. In VAD, the main goal is to distinguish segments of a signal that contain speech from those that do not, so that any processing effort may be focused only on information consisted of speech. In SSL, the main idea is to explore the spatial information of the acoustic signals through microphone array beamforming techniques, to enhance the speech of a source of interest while suppressing those of competing sources and lowering environment noise [Brandstein and Ward 2001].

In most existing works, VAD and SSL are approached for a single speech source case, what might not be appropriate for a number of situations. In applications such as multi-conferences, gaming scenarios, automatic information retrieval, and also HCI, it is often desirable to distinguish between different users that might overlap their speeches. This ends up extending both VAD and SSL to more complex problems than in cases where a single speaker is considered. Our methods, however, are able to address such adverse situations by using joint audio-video (multimodal) signal processing. Our original work (Master's Dissertation) has presented a collection of previously produced articles that have been published [da Silveira et al. 2010, Blauth et al. 2012, Minotto et al. 2012, Minotto et al. 2013] or that are already accepted for publication [Minotto et al. 2014].

The referred works have dealt with VAD and SSL of one and multiple speakers in different ways. We have explored distinct methods for fusing audio and video data, achieving, in all works, above 90% accuracy for VAD, and below 11 cm error for three-dimensional SSL. Additionally, due to the necessity for real-time processing of HCI applications, we have also developed an efficient GPU version of the Steered Response Power with Phase Transform (SRP-PHAT), which is a key audio processing technique used in our algorithms [Minotto et al. 2012].

At this point, it is important to illustrate the employed data capture system in our works. Figure 1 shows a photo and a schematic representation of it. A linear microphone array and a common color camera are used, from which multi-channel sound and VGA images are acquired, respectively, and then used as primary inputs for our algorithms. It is important to mention that from this setup, a constraint arises: the users must be facing the capture hardware. We consider this to be a reasonable assumption, since it is typically the case of HCI applications.



(a) (b)

Figure 1: (a) Scheme of our prototype system. (b) Photo of the capture hardware.

The remainder of this document is organized as follows. Section 2 summarizes the contributions of our work. In Section 3 we present some of the results obtained in our most recent developed technique, and final considerations and acknowledgments are made in Section 4.

2. Contributions of our Work

In general, our work reflects as contributions to various research fields. The produced methods deal with VAD and SSL of single and multiple speech sources in a typical HCI-like acoustic scenario. For such, we have integrated several audio and video signal processing techniques through different multimodal fusion approaches and machine learning techniques. A detailed description of each work's contribution is presented next.

In [Blauth et al. 2012], we presented an approach that performs single speaker VAD using audio information only (video is included for SSL). We have developed a Hidden Markov Model (HMM) competition scheme, through which VAD is performed by analyzing the output of the SRP-PHAT microphone array beamforming technique. The SRP-PHAT is mainly an SSL algorithm, and is known to be robust in realistic conditions. We extend it to a VAD method by assessing the spatio-temporal behavior of the dominating sound source against two HMMs, one that models speech situations and one that

models silence. The dominant speaker is classified as active or inactive by comparing the likelihood of the spatial features generated from the HMMs. For the SSL step, video cues are included from the results of a face-tracking algorithm [Bins et al. 2009], by weighting the SRP-PHAT result based on the tracked faces positions. This technique has shown robustness when applied in realistic conditions, for it achieved average accuracies of 92% for VAD and 96.3% for SSL even under situations with purposely generated noise.

The mentioned work is then expanded to a multimodal technique, in [Minotto et al. 2013]. We combine our audio-based approach [Blauth et al. 2012] to the video-based approach of [Lopes et al. 2011] through a decision fusion scheme. We study many supervised classification algorithms for merging the results of the individual unimodal classifiers. The well known Machine Learning software Weka [Hall et al. 2009] is used for exploring a variety of approaches, through which it is concluded that a C4.5 decision tree [Quinlan 1993] presents the best benefits in this scenario (trade-off between accuracy and speed, besides also being robust against overfitting). Our results suggest the proposed features are stable enough to suit many classification algorithms, outputting a VAD accuracy above 93%. As another contribution of the work, we also analyze the robustness of our approach to adverse situations (intentionally generated), confirming that one modality in fact properly compensates for the other's flaws. Therefore, besides increasing the overall accuracy of the previous technique, this multimodal approach also provides stability to cases where one of the data streams becomes unreliable.

In [Minotto et al. 2014], a multimodal approach for simultaneous speakers VAD and SSL is developed, by extending the ideas from both previously mentioned papers. The HMM competition scheme is altered in order to deal with multiple speakers cases. An optical-flow algorithm is used to assess lip movements of each participant, which generates visual features as inputs to a Support Vector Machine (SVM) classifier. The SVM outputs a video-based VAD probability for each potential speaker, and the audio modality is processed by the multi-user HMM competition scheme, at which point the video probability is incorporated. This characterizes the combination of both modalities, and is considered a mid-fusion approach. The final VAD decision is performed by the analysis of the competing models, and its results are also reused for generating a final SSL position (recalling the HMM scheme uses the SRP-PHAT SSL method). An average VAD accuracy of 95.06% is obtained for up to three simultaneous speech sources, and three-dimensional SSL is performed for the active speakers with an average distance error of 10 cm between the estimated position to the true positions.

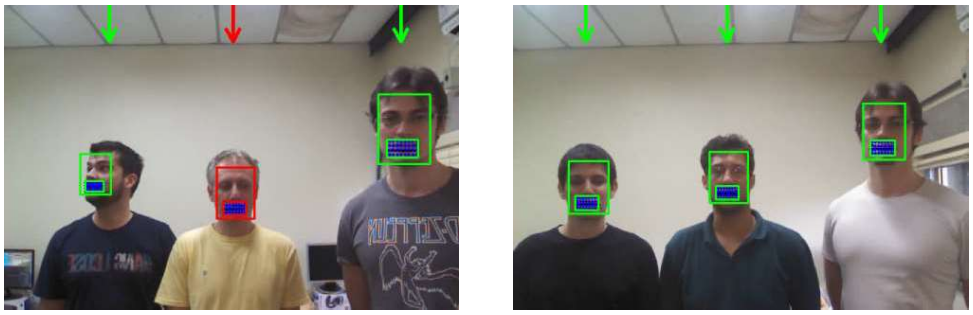
Finally, the work in [Minotto et al. 2012] describes the GPU implementations developed to achieve real-time processing in our multimodal systems. As it may be observed, the aforementioned approaches employ the SRP-PHAT algorithm through a microphone array. Despite its known robustness, the SRP-PHAT has a high computational cost as a drawback. For this reason, we have implemented two Compute Unified Device Architecture (CUDA) versions of the algorithm, as well as one for the Cubic Splines Interpolation, which is commonly applied as a part of the SRP-PHAT itself. Using such methods we are able to achieve real-time processing in our previously mentioned VAD/SSL works, which is a necessity of most HCI-related application.

From the referred papers we may notice the accomplished progression of our techniques, from a single-speaker unimodal work to a multiple speakers multimodal VAD and

SSL one. Incorporating extra modalities of data into our methods was inherent, given the more realistic the scenario is, the more complex the problem becomes. While other works also tackle such problems using audiovisual approaches, most part resort to some impractical capture system (as a 500-element microphone array) or deal with the single-speaker case only [Thiran et al. 2010]. Furthermore, there exist few to none multimodal public data-sets for performing tests and benchmarks. For this reason we made available all our recorded sequences. Each of the referred works has a link their corresponding used sets.

3. Overall Results

Since the entirety of our work comprises different techniques, this section will focus on presenting the results related to the work published most recently [Minotto et al. 2014], which we consider to be the one that tackles an acoustic scenario more challenging than the others. Figure 2 illustrates our algorithm running for two different multimodal sequences. The color of the bounding boxes indicates speech (green) or silence (red) detected by our algorithm, while the arrows indicate the true VAD status of the users (manually created ground-truths). Intuitively, matching colors mean our algorithm performed a correct classification. The blue dots within the mouth bounding box illustrate the points tracked by the Lucas-Kanade optical-flow algorithm, which is responsible for outputting a visual feature to the video-based module of our system. The experimental results presented next are referred to situations similar to those of the Figs. 1(a) and 1(b), which are extracted from the sequences available in our data-set.



(a) (b)

Figure 2: (a) Situation with two speakers active. (b) All users speaking at the same time.

In more details, the data-set consists of several multimodal sequences ranging from 40 to 60 seconds of duration each[1]. Among those, we present results for scenarios containing two and three speakers, which are named `Two1` to `Two5`, and `Three1` to `Three5`, respectively. In all recordings, the users randomly chat in Portuguese, alternating between speech and silence moments, and intentionally overlapping their voices at times. Sequences with two users consist of sections of individual speech (implying in individual silence of the other speaker), simultaneous speech, and simultaneous silence. For the sequences with three users, the same procedure is used, but applied to all possible combinations of speakers. Furthermore, all sequences in the data-set contain some sort of natural noise, such as people talking in the background, air-conditioning working, door slams, and the fans from other computers.

[1]Details on the setup may be found in `http://www.inf.ufrgs.br/~crjung/MVAD-data/mvadsimult.htm`, where the multimodal sequences with ground truth can be downloaded.

For measuring the VAD accuracy, we have manually labeled each speaker at each frame as active or inactive, and ran three experiments for each sequence (which are shown in Table 1): one for our multimodal approach (named "Ours"), one for the VAD work of [Sohn et al. 1999] (called "Sohn"), and a third for the VAD module of the G729B codec [ITU-T 1996]. Since these competitive algorithms were designed for single speaker scenarios, we chose the left-most speaker as the reference one, so that a fair comparison is made. More specifically, the obtained results of each experiment are compared to the ground truth of the left person only. As expected, both competitive approaches perform worse for simultaneous speakers, owing to the speech of other users acting as noise to the speech of the considered one. This demonstrates the importance of multi-speaker VAD that is able to properly isolate different users, as is the case of our method.

Table 1: Comparison of VAD methods.

Sequence	Ours	Sohn	G729B
Two1	96.89%	78.65%	78.91%
Two2	97.81%	79.85%	79.33%
Two3	95.93%	80.95%	70.48%
Two4	96.71%	81.71%	75.55%
Two5	94.11%	82.66%	78.75%
Avg. (Two)	96.29%	80.76%	76.60%
Three1	94.02%	75.29%	63.42%
Three2	92.39%	80.74%	78.98%
Three3	94.94%	82.48%	74.13%
Three4	92.10%	74.03%	75.43%
Three5	95.58%	79.74%	69.14%
Avg. (Three)	93.81%	78.46%	72.22%

Table 2: SSL accuracy in terms of Euclidean distance error, in meters

Sequence	Audio	Video	Multimodal
Two1	0,1479	0,1826	0,1275
Two2	0,1365	0,0963	0,0950
Two3	0,1422	0,1358	0,0999
Two4	0,2033	0,1522	0,1020
Two5	0,1381	0,1278	0,1148
Three1	0,2007	0,1643	0,1096
Three2	0,1887	0,1311	0,1172
Three3	0,2040	0,1218	0,1022
Three4	0,1976	0,1338	0,1052
Three5	0,1845	0,1298	0,1102
Average	**0,1743**	**0,1375**	**0,1084**

To evaluate the SSL performance of our algorithm, we have computed the Euclidean distance between the found locations and the labeled locations as error measures (results are shown in Table 2). This process was repeated for the video and audio modalities alone as well as for our multimodal SSL approach. For the video modality, an inverse projective mapping from the image plane coordinates to 3D world ones has been used (using the face scale to estimate the depth). For the audio modality, the SRP-PHAT was directly applied, and for the multimodal one, the previously summarized multi-user HMM-based fusion has been used (a more detailed description may be found in [Minotto et al. 2014]). It may be observed that our SSL approach presents accuracy gains over the audio and video modalities alone. An average error of 10.84 cm exists when estimating the speakers' 3D position, which is about twice the average length of the human mouth, meaning it is precise to the point no speaker is confused as being another.

4. Final Considerations

This document summarized our dissertation in the field of Voice Activity Detection and Sound Source Localization. We presented the chronological progress of a single-speaker unimodal technique to a more complex multimodal multi-speaker one, by outlining the published articles resulted from this work. Our techniques focused on realistic environments, which were exposed to high levels of noise, and under the constraint of real-time processing. The next step to further improve our simultaneous speaker multimodal approach is to include other modalities of data. Experiments using information from a RGB-D camera have already been conducted, achieving VAD accuracies above 95%.

Acknowledgments of this work go to Hewlett-Packard (HP) Brasil Ltda., for financing our work, using incentives of Brazilian Informatics Law (Law n 8.2.48 of 1991). Furthermore, all source code produced is property of HP, and has been incorporated into the company's C++ media processing framework as VAD and SSL modules.

References

Bins, J., Jung, C. R., Dihl, L., and Said, A. (2009). Feature-based face tracking for videoconferencing applications. In *Multimedia, 2009. ISM '09. 11th IEEE International Symposium on*, pages 227 –234.

Blauth, D. A., Minotto, V. P., Jung, C. R., Lee, B., and Kalker, T. (2012). Voice activity detection and speaker localization using audiovisual cues. *Pattern Recognition Letters*, 33(4):373 – 380.

Brandstein, M. and Ward, D. (2001). *Microphone arrays: signal processing techniques and applications*. Digital signal processing. Springer.

da Silveira, L. G., Minotto, V. P., Jung, C. R., and Lee, B. (2010). A GPU Implementation of the SRP-PHAT Sound Source Localization Algorithm. *The 12th International Workshop on Acoustic Echo and Noise Control*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

ITU-T (1996). A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70, Annex B.

Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2):116–134.

Lopes, C., Goncalves, A., Scharcanski, J., and Jung, C. (2011). Color-based lips extraction applied to voice activity detection. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1057 –1060.

Minotto, V., Jung, C., da Silveira, L., and Lee, B. (2012). GPU-based Approaches for Real-Time Sound Source Localization using the SRP-PHAT Algorithm. *International Journal of High Performance Computing Applications*.

Minotto, V., Jung, C., and Lee, B. (2014). Simultaneous-speaker voice activity detection and localization using mid-fusion of svm and hmms. In *IEEE Transactions on Multimedia. Accepted por publication. Available at IEEE EARLY ACCESS ARTICLES.*

Minotto, V., Lopes, C., Scharcanski, J., Jung, C., and Lee, B. (2013). Audiovisual voice activity detection based on microphone arrays and color information. *Selected Topics in Signal Processing, IEEE Journal of*, 7(1):147–156.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Sohn, J., Member, S., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Process. Lett*, 6:1–3.

Thiran, J.-P., Marqués, F., and Bourlard, H. (2010). *Multimodal Signal Processing, Theory and Applications for Human-Computer Interaction*. Academic Press.