IDIAP RESEARCH REPORT

AUDIO SPATIO-TEMPORAL FINGERPRINTS FOR CLOUDLESS REAL-TIME HANDS-FREE DIARIZATION ON MOBILE DEVICES

Danil Korchagin

Idiap-RR-08-2011

MAY 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

AUDIO SPATIO-TEMPORAL FINGERPRINTS FOR CLOUDLESS REAL-TIME HANDS-FREE DIARIZATION ON MOBILE DEVICES

Danil Korchagin

Idiap Research Institute
Martigny, Switzerland

ABSTRACT

In this paper, we propose a new low bit rate representation of a sound field and a new method for the corresponding cloudless low delay hands-free diarization suitable for low-performance mobile devices, e.g. mobile phones. The proposed audio spatio-temporal fingerprint representation results in low bit rate (500 bytes/second), however contains complete information about continuous audio tracking of multiple acoustic sources in an open, unconstrained environment. The core of the algorithm is based on simultaneous multiple data stream processing using audio spatio-temporal fingerprint representation to cover higher level events relevant for diarization, e.g. turns, interruptions, crosstalk, speech and non-speech segments. Performance levels achieved to date on 5 hours of hand-labelled datasets have shown the feasibility of the approach at the same time as resulting in 7.58% CPU load on 1-core ultra-low-power mobile processor running at 1 GHz and low algorithmic delay of 112 ms.

Index Terms **—** Microphone arrays, array signal processing, mobile computing, source coding

1. INTRODUCTION

Speaker diarization is the task of determining "who spoke when" in an audio stream. The diarization systems identify the speech segments corresponding to each speaker and estimate the number of speakers. Conventional speaker diarization systems [1] use an ergodic Hidden Markov Model (HMM) with speakers as HMM states. Good results were achieved by the systems using the combination of Mel-Frequency Cepstral Coefficients (MFCC) and Time Difference of Arrival (TDOA) features [2] with arrays composed of different number of microphones, while performance of standalone TDOA features was estimated as poor in respect to MFCC [3]. TDOA features can be used without prior knowledge of the geometry of the microphone array. In the case that the geometry of the microphone array is known in advance, TDOA features can be replaced by the speaker locations, which are often used as complementary features to conventional MFCC [4].

Figure 1. Conceptual family environment setup.

In our work we rely on prior knowledge of the geometry of the microphone array and perform the study on the feasibility to achieve reasonable performance with focus on computationally efficient multi-source localisation as the primary standalone features for the speaker diarization system in an open, unconstrained environment.

Typically, speaker localisation can either be done in the audio modality, video modality or multimodality. The first one implies a microphone array usage, while the second one is based on movement detection. Multimodal localisation allows results to be less affected by noise in the audio modality, although it increases significantly the CPU load. The general multimodal approach is to transform the data in such a way that a correlation between the audio and a specific location in the video is found [5, 6]. Other multimodal techniques use score-level fusion via estimation of the mutual information between the average acoustic energy and the pixel value [7], probability density estimation [8] or a trained joint probability density function [9]. While in our study we consider the audio modality only, an extension to multimodal techniques is applicable.

Finally, we present an overview of a few potential application systems, where the proposed method can be exploited. Nowadays most of the applications of speaker diarization/localisation systems are restricted by business area because of its complexity and cost aspects. Though it could happen that one day in the future the corresponding system setup can be as simple as randomly placing a mobile phone on a table (Figure 1).

2. COMPUTATIONALLY EFFICIENT DIARIZATION

To achieve seamless low delay real-time performance the algorithm presented in this paper was implemented and evaluated as a plug-in for the data-flow architecture Tracter [10]. Data-flow is a well established signal processing technique that represents individual processing elements (plug-ins) as vertices in a directed graph. The data is propagated through the graph using a "pull" mechanism, instigated by the sink. The pull mechanism also allows the dataflow to be driven by the Weighted Finite State Transducer (WFST) decoder [11], if required by a subsequent application.

The core capture device for the system is any type of embedded (not yet available on the market) or external microphone array, e.g. an audio diamond array with four omnidirectional microphones or USB-based Microcone [12] (Figure 2). Audio signals from the microphone array are retrieved in real-time and contain interleaved 4+ channel PCM audio in 16-bit samples at 16 kHz.

Figure 2. External diamond array based on 4 omnidirectional AKG C562CM microphones (above), Beyerdynamic MPC23SW (middle) and USB-based Microcone (below).

2.1. Instantaneous spatial fingerprints

We propose to define instantaneous spatial fingerprints as bit patterns of overlapping sector-based acoustic activity measures, where each sector is represented by 1 bit of information (Figure 3). The corresponding instances in time refer to processing frames of 32 ms length.

Each sector is defined as a 36° wide and 60° high (from the horizontal plane) connected volume of physical space around the microphone array. The sectors are taken in the horizontal plane in steps of 6°. This results in a total of 60 sectors. Wider sectors in smaller steps allow to avoid jittering of acoustic directions and smooth acoustic tracking of dynamic sources.

The sector activity measure [13] is defined as integrated within the sector point-based steered response power with phase transform weighting (SRP-PHAT). SRP-PHAT [14] in turn is defined as the sum of generalized cross correlations with phase transform weighting (GCC-PHAT [15]) for each microphone pair. Further, a sparsity assumption is applied for each frequency bin via minimisation of phase error and the sector activity measures are normalised by the volume of the sector. The sector activity measure relies only on the geometry of the microphone array and does not depend on prior knowledge of the room dimensions.

Each sector activity measure is thresholded to keep a binary decision, which gives us 60 bits of data for 360° spatial representation per each instance in time. This information is stored as one 64 bit integer value.

Finally, the spatial fingerprint is multiplied by the predefined mask. This multiplication results in directional filtering of the predefined areas of interest, elimination of unnecessary post calculations and outlier removal. It can be very helpful in the case of interconnected environments, where audiovisual links do not have an echo suppression mechanism. For example, remote parties can be shown on a TV screen (Figure 1), while the corresponding TV zone is out-of-interest for local diarization (distributed diarization can be driven as the superposition of local diarizations from all interconnected environments).

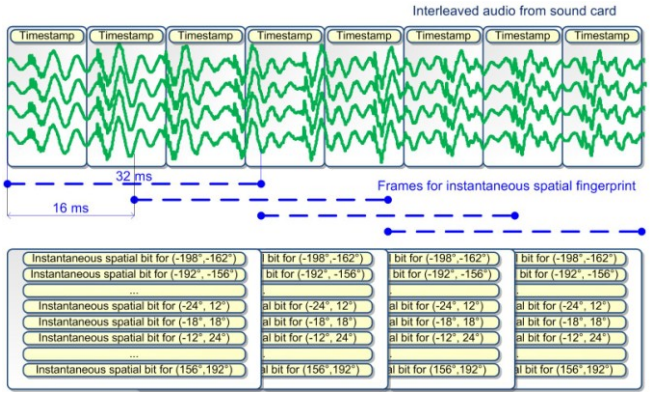

Figure 3. Slicing frames for spatial fingerprints.

2.2. Spatio-temporal fingerprint processing

We define the spatio-temporal fingerprint representation as an array of temporally connected spatial fingerprints taken in steps of 16 ms. This results in a 2D bit pattern (Figure 4) with a total of 62.5 columns per second and the low bit rate of 500 bytes/second (62.5 long integer values of 64 bits each). The spatio-temporal fingerprints are defined as subsets of the spatio-temporal fingerprint representation. The length of spatio-temporal fingerprints depends on the application and can vary from 32 ms to several seconds.

The intersection fingerprint is defined as an intersection in the time domain of all elements within a spatio-temporal fingerprint. Similarly, the union fingerprint is defined as a union in the time domain of all elements within a spatio-temporal fingerprint. While the default length for all spatio-temporal fingerprints has been chosen to be 112 ms, in the next section we present the results of a study for shorter and longer spatio-temporal fingerprints as well. The resulting intersection and union fingerprints are normalised at each time instance by keeping one centralised bit per active source.

Due to the compact fingerprint representation we can benefit from the exploitation of multiple (60) data streams against a single instruction stream to perform operations, which may be naturally parallelized. This approach is widely used in many areas of Information and Communication Technologies (ICT) nowadays and is often referred as Single Instruction, Multiple Data (SIMD) streams according to Flynn's taxonomy [16]. In our study the SIMD approach is exploited for most of the bitwise operations (e.g., intersection, union and normalisation operations are represented via bitwise AND, OR and XOR operators).

Because the speech is known to be intermittent, we introduce at the next step a State Transition Network (STN, Figure 5), which allows us to enforce speech continuity. The intersection and union fingerprints are used for transitions from one state to another, while only intersection fingerprints are used for continuous tracking of acoustic sources within “speech” and “crosstalk” states.

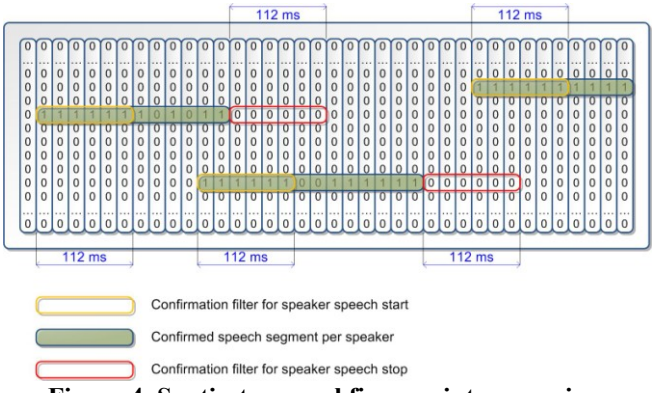

Figure 4. Spatio-temporal fingerprint processing.

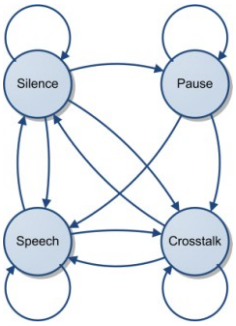

Figure 5. State transition network topology.

To allow low delay real-time and seamless diarization, transitions between states are associated with events. These events include, but are not limited to: crosstalk, successful and unsuccessful interruptions, turn taken, voice activity, no voice activity and pause. The pause differs from no voice activity by a longer confirmation time filter and has an impact only on an orchestrated video conferencing system, presented in the following section.

The transition into a state with a lower number of acoustic sources is performed based on a union fingerprint, while the transition into a state with a higher number of acoustic sources is performed based on an intersection fingerprint. The number of simultaneous acoustic sources is estimated as the Hamming distance between the last confirmed fingerprint and 0. The corresponding spatial locations of the active sources are computed as bit positions inside the confirmed intersection fingerprint multiplied by 6°.

The turn taken event is used as an additional trigger for speech segmenting to allow better association with the respective source. Taking into account that acoustic sources are not static in general, we cannot apply the Hamming distance between two intersection fingerprints to establish the turn taken event. The turn taken is confirmed only in the case that a shift of the active source bit position from the previous confirmed state is higher than the predefined threshold. Otherwise the state is updated with the new location without issuing the turn taken event. If the speech segments of different speakers overlap or concatenate each other, the turn taken event is augmented by detection of successful or unsuccessful interruption, which could be employed by subsequent social signal processing. The turn taken event can also be used for estimation of number of speakers over a predefined time window.

The crosstalk event is triggered if there is concurrent speech from two or more sources for at least 112 ms. Depending on the application, the meaning of the event is different. For example, in the case of an orchestrated video conferencing system, it would mean switching to a wider shot to cover crosstalking participants. In the case of a subsequent speech transcription, this event could trigger corresponding reinitialization of associated beamformers for better source separation.

3. RESULTS AND EVALUATIONS

The experiments were performed on real life hand-labelled datasets (3h 50min for dataset 1 (DS1) with echo suppression enabled [17]; 1h 20min for dataset 2 (DS2) [18] with echo suppression disabled, lower SNR and higher density of people). The datasets contain recorded gaming sessions with video chat enabled and follow the systematic data description presented in [18]. Each room was recorded and analysed separately and contained 2 people for dataset 1 and up to 4 people for dataset 2 (Figure 6). The microphone array configuration consists of four omnidirectional Beyerdynamic MPC23SW microphones (Figure 2, middle) for dataset 1 and AKG C562CM (Figure 2, above) for dataset 2. The microphones are arranged on the corners of a square. The distance between two microphones on opposite corners, i.e. the diagonal of the square, was chosen as 4.5 cm for the Beyerdynamic MPC23SW and 3.2 cm for the AKG C562CM. The second part of the TA2 database [18], recorded with a circular array of eight omnidirectional microphones, was excluded from our consideration due to non-compact size of the microphone array (the diagonal is 20 cm instead of 3.2-4.5 cm).

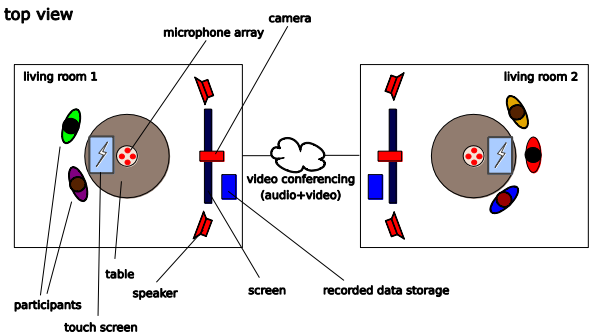

Figure 6. Evaluation setup [18].

The sector of interest for directional filtering was defined as [-110°, 110°] with respect to the reference direction of 0°, defined as an imaginary arrow intersecting the camera and the centre of the microphone array, facing the participants. This allows us to eliminate remote parties in case of disabled echo suppression (DS2).

The dependency between precision and recall values is estimated via application of different thresholds at the step of packing data into the spatio-temporal fingerprint representation. In Figure 7 this dependency is illustrated for speech/non-speech detection and speaker match. Precision is defined as the accumulated length of true positive segments (segments correctly detected as belonging to the positive class) divided by the total length of segments detected as belonging to the positive class (the sum of true positive and false positive segments). Recall is defined as the accumulated length of true positive segments divided by the total length that actually belongs to the positive class (the sum of true positive and false negative segments).

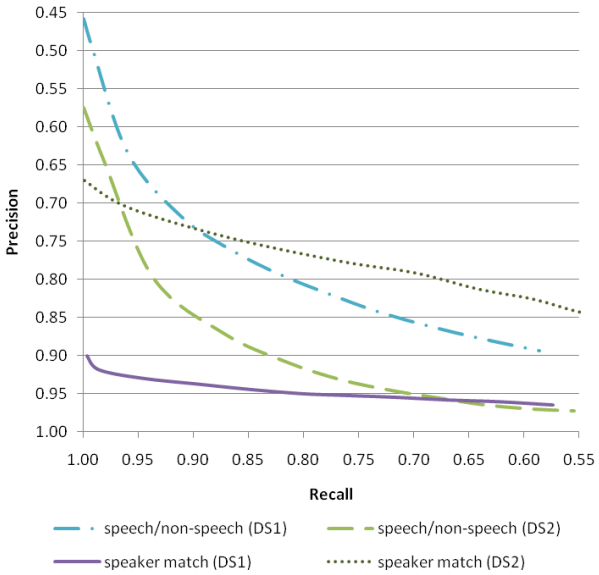

Figure 7. Precision versus recall for speech/non-speech detection and speaker match (DS1+DS2).

It is clearly visible, that for dataset 1 the speaker match (solid line) shows better performance than the speech/non-speech detection (dash dot line). Higher precision values correspond to lower recall values and vice-versa. Although only dataset 1 is echo-cancelled we were able to achieve good precision/recall levels for the speech/non-speech detection on dataset 2 due to application of the directional filtering, nevertheless the higher people density in dataset 2 resulted in lower precision/recall values for the speaker match. Depending on the subsequent application, the precision and recall priorities can be different.

Another parameter, which has a strong impact on the performance of the system, is the length of spatio-temporal fingerprints. This parameter defines as well the algorithmic delay of the proposed approach. In Figure 8 we illustrate how the length of spatio-temporal fingerprints impacts the precision and recall values for speech/non-speech detection and speaker match. The best results were achieved for spatio-temporal fingerprints within [112 ms, 192 ms].

We were able to achieve 95.0% precision and 80.3% recall for the speaker match with a delay of 112 ms for dataset 1 (77.7% precision and 76.7% recall for dataset 2). For the same parameter set, the speech/non-speech detection resulted in 79.4% precision and 82.1% recall for dataset 1 (87.1% precision and 87.2% recall for dataset 2). While very low bit rate representation has no impact on precision/recall levels for speaker match (only the position of the sector is used regardless how many bits are dedicated to the sector representation), it can bring additional speech/non-speech detection errors into the system (estimated precision/recall levels for standard energy based speech/non-speech detection are 89.2% and 88.4% for dataset 1, 72.7% and 71.5% for dataset 2).

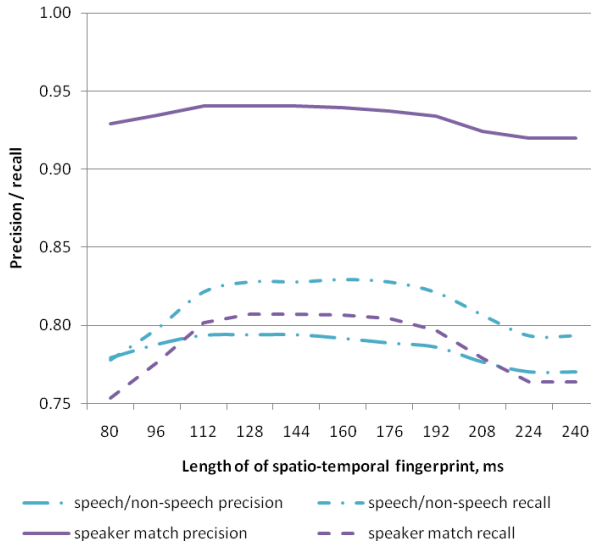

Figure 8. Precision/recall versus fingerprint length for speech/non-speech detection and speaker match (DS1).

Known meeting-wise speaker error rates for CPU-intensive state of the art techniques [2] are as low as 7.0% for realigned MFCC+TDOA combination of the HMM/GMM system with optimal weights and for Kullback-Leibler based realigned MFCC+TDOA combination of the information bottleneck system with optimal weights. In the case of automatic weights, overall speaker error rates are 13.6% and 9.9% correspondingly. These state of the art estimates are given only as an overview and cannot be used for direct comparison with the proposed method as the data, hardware and scenario used in our experiments differ (slightly) from the data, hardware and scenario used in [2]. In addition the state of the art systems have a delay of 500 ms and a state of minimum 3 seconds duration, while we were able to achieve reasonably good results with algorithmic delay and minimum state duration as low as 112 ms. We should note that the algorithmic delay does not include capturing delay, which in turn can result in additional 10-20 ms.

CPU load (in respect to a single CISC core running at 1 GHz; averaged result for a session of 3h 50m) for the proposed approach was 7.58%, most of it due to sound field packing into spatio-temporal fingerprints. The complete audio processing chain including feature extraction components for subsequent ASR resulted in 15.94% CPU load (ASR decoder based on Weighted Finite State Transducer [11] was omitted from this test because it is the most CPU-intensive task, not suitable for mobile devices). Thus we can conclude that the computational efficiency of the proposed approach is suitable for mobile devices and results to date give us good prerequisites for future research in spatio-temporal fingerprint processing to be able to achieve higher precision/recall values and lower CPU load.

4. APPLICATION SYSTEMS

In this section we describe a few potential application systems that could benefit from using the proposed technique in real life. To our knowledge there is no single mobile phone available on the market with an integrated microphone array designed for distant use; nevertheless this can be resolved by using any of the USB microphone arrays available on the market (e.g., Microcone [12]). In the future, we presume, the situation can be changed and compact microphone arrays can be directly integrated by manufactures into next generation mobile phones.

One of the potential systems is an orchestrated video conferencing system with spatially separated non-intrusive sensors. By placing the sensors at their individually optimal locations, better performance of semantic information extraction can be expected (as opposed to other systems [19, 20], relying on collocated sensors). Semantic information, extracted on the fly, is used to produce an orchestrated video chat [21] by taking pure video streams from multiple cameras and at each point in time choosing the perspective that best represents the social interaction. While for many of us, video conferencing systems are still associated with expensive business solutions from Tandberg/Cisco [22] or Polycom [23], recently there were several attempts to enter the home entertainment market by leading video conferencing companies and research projects.

The TA2 project (Together Anywhere, Together Anytime [24]) is a large scale research project, which tries to understand how corresponding technologies can help to nurture family-to-family relationships to overcome distance and time barriers in home environments. The technique presented in this paper can potentially decrease the complexity of the complete system by execution of corresponding bits of the scene analysis within a mobile phone. Further it can directly communicate corresponding acoustic events with a low delay via wireless interface to the orchestrated video conferencing system by the same principle as was described in our previous work [25]. In future work we are going to exploit the proposed method within a distributed multimodal analysis system to be employed by the orchestrated video conferencing system.

Another potential system for exploitation is automatic multiparty speech transcription, which allows significantly improved semantic value of the media data. The corresponding technologies have already entered the market as cloud-based services, provided by Koemei [26], Google [27], Nuance [28] and others. The last bridge to boost these cloud-services to everyone can rely on a computationally efficient multiparty capturing/segmenting method (e.g., as described in this paper) running within a simple mobile phone, placed randomly on a table as illustrated in Figure 1. Especially it could have big success in the home entertainment market, where the price of a solution plays a significant role in its exploitation.

5. CONCLUSIONS

We have shown the feasibility of using audio spatio-temporal fingerprints as a computationally efficient solution for low delay hands-free diarization, suitable for low-performance mobile devices. Performance levels achieved to date on 5 hours of hand-labelled datasets have shown sufficient reliability at the same time as fulfilling real-time processing requirements with an algorithmic delay of 112 ms and a sound field bit rate of 500 bytes/second. The estimated CPU load is 7.58% on a 1-core ultra-low-power mobile processor running at 1 GHz. An overview of potential application systems shows that there is a demand for low cost computationally efficient solutions. The results are promising for future research in audio spatio-temporal fingerprint processing in respect of accuracy gap minimisation between the proposed computationally efficient method and CPU-intensive state of the art algorithms.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project “Together Anywhere, Together Anytime” (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793. We are grateful to reviewers for their valuable feedback and Philip N. Garner for his valuable help at various stages of this work.

7. REFERENCES

[1] Ajmera, J., "Robust audio segmentation", *Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL)*, 2004.

[2] Vijayasenan, D., Valente, F. and Bourlard, H., "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization", *in IEEE Trans. on Audio Speech and Language Processing*, 19(2), 2011.

[3] Pardo, J., Anguera, X. and Wooters, C., "Speaker Diarization for Multi-Microphone Meetings Using Only Between-Channel Differences", *in Proc. of MLMI*, Bethesda, USA, 2006.

[4] Lathoud, G. and McCowan, I., "Location Based Speaker Segmentation", *in Proc. of ICASSP*, Hong Kong, China, 2003.

[5] Slaney, M. and Covell, M., "Facesync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks", *in Proc. of Neural Information Processing Systems*, pp. 814-820, 2000.

[6] Monaci, G., Escoda, O. D. and Vandergheynst, P., "Analysis of Multimodal Sequences Using Geometric Video Representations", *in Signal Processing*, vol. 86, pp. 3534-3548, 2006.

[7] Hershey, J. and Movellan, J., "Audio Vision: Using Audio-Visual Synchrony to Locate Sounds", *in Proc. of Neural Information Processing Systems*, pp. 813-819, 1999.

[8] Nock, H., Iyengar, G. and Neti, C., "Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study", *in Proc. of CIVR*, Urbana-Champaign, USA, 2003.

[9] Gurban, M. and Thiran, J., "Multimodal Speaker Localization in a Probabilistic Framework", *in Proc. of EUSIPCO*, Florence, Italy, 2006.

[10] Garner, P. N., Dines, J., "Tracter: A Lightweight Dataflow Framework", *in Proc. of Interspeech*, Makuhari, Japan, 2010.

[11] Garner, P. N. et al., "Real-Time ASR from Meetings", *in Proc. of Interspeech*, pp. 2119-2122, Brighton, UK, 2009.

[12] Dev-Audio, "Microcone: The Intelligent Microphone Solution for Recording Group Meetings", http://www.dev-audio.com, 2010.

[13] Lathoud, G. and McCowan, I. A., "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays", *in Proc. of SAPA*, Jeju, Korea, 2004.

[14] Knapp, C. H. and Carter, G. C., "The generalized correlation method for estimation of time delay", *in IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24(4), pp. 320–327, 1976.

[15] DiBiase, J., Silverman, H., and Brandstein, M., "Robust localization in reverberant rooms", *in Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springe, ch. 8, 2001.

[16] Flynn, M., "Some Computer Organizations and Their Effectiveness", *in IEEE Trans. Comput.*, C-21: 948, 1972.

[17] Kuech, F. et al., "Acoustic Echo Suppression Based on Separation of Stationary and Non-Stationary Echo Components", *in Proc. of IWAENC*, Seattle, USA, 2008.

[18] Duffner, S., Motlicek, P. and Korchagin, D., "The TA2 Database: A Multi-Modal Database from Home Entertainment", *in Proc. of ICSAP*, Singapore, 2011.

[19] Bohus, D. and Horvitz, E., "Dialog in the Open World: Platform and Applications", *in Proc. of ICMI*, Cambridge, USA, 2009.

[20] Otsuka, K. et al., "A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization", *in Proc. of ICMI*, Chania, Greece, 2008.

[21] Stevens, T. et al., "Enhancing Social Communication Between Groups", *in Proc. of ICIN*, Berlin, Germany, 2010.

[22] Cisco, Tandberg department, "Video Conferencing Solutions and Telepresence Products", http://www.tandberg.com, 2011.

[23] Polycom, "High Definition Telepresence Systems", http://www.polycom.com, 2011.

[24] Integrating Project within the European Research Programme 7, "Together Anywhere, Together Anytime", http://www.ta2-project.eu, 2008.

[25] Korchagin, D. et al., "Hands Free Audio Analysis from Home Entertainment", *in Proc. of Interspeech*, Makuhari, Japan, 2010.

[26] Koemei, "Cloud-Based Speech Recognition Solution for Multiparty Conversations", http://www.koemei.com, 2010.

[27] Google, "Cloud-Based Speech for Mobile and the Web", *in SpeechTEK Europe*, London, UK, 2011.

[28] Nuance, "Nuance on Demand: Cloud-Based Voice Platform", http://www.nuance.com, 2010.