Hands Free Audio Analysis from Home Entertainment

Danil Korchagin, Philip N. Garner, Petr Motlicek

Idiap Research Institute, Martigny, Switzerland

{Danil.Korchagin,Phil.Garner,Petr.Motlicek}@idiap.ch

Abstract

In this paper, we describe a system developed for hands free audio analysis for a living room environment. It comprises detection and localisation of the verbal and paralinguistic events, which can augment the behaviour of virtual director and improve the overall experience of interactions between spatially separated families and friends. The results show good performance in reverberant environments and fulfil real-time requirements.

Index Terms: real-time audio processing, direction of arrival, speech meta-data

1. Introduction

The TA2 project (Together Anywhere, Together Anytime) [1] seeks how technology can help to nurture family-to-family relationships to break down distance and time barriers. This is something that current technology does not address well: modern media and communications serve individuals best, with phones, computers and electronic devices tending to be user centric and providing individual experiences. In this sense, we are interested in effective hands free audio analysis system to be employed by virtual director to make communication and engagement easier among groups of people separated in space and time.

The system includes verbal and paralinguistic event detection, localisation, association and fusion. This involves far-field processing of voice activity, estimation directions of arrival, Automatic Speech Recognition (ASR) with keywords and proper names spotting. Although the accuracy of far-field ASR is not yet good enough to be exploited for accurate real-time transcription, it is still suitable to augment the behaviour of virtual director. Words in the transcript are used to search for participant proper names relevant to the group of people or keywords relevant to the enabled scenario. Further, a virtual director takes these events into account with other cues coming from the game engine, aesthetic rules, maximum frequency of cutting, and makes decisions on the selection and transformation of audiovisual content.

The system for hands free audio analysis relies on external interleaved multichannel audio from multiple distant microphones for online data processing. Direction of arrival (who spoke when) is integrated into main processing chain. This information is used to mark up the voice activity and keywords / proper names assigning them to speaker locations. In addition, the confidence of this association is estimated, which helps to virtual director to choose most appropriate shots for cases of confident and non-confident associations.

In the context, TA2 presents several challenges: results are supposed to be low delay, online and real-time to allow just-in-time reasoning on audiovisual streams. Further the results are supposed to be localised in space and synchronised with external time clock.

2. A real-time architecture

Separately real-time voice activity detector, automatic speech recogniser, keyword spotter and direction of arrival estimator are by no means a new concept. They have been in use in research and commercial products for many years. The novelty here comes from the low delay results association, confidence estimation and keeping the results synchronised with remote time clock of external entertainment server.

The system architecture is built around several modules comprising a large vocabulary ASR decoder known as Juicer and real-time framework known as Tracter [2]. Both input and output of the system are non-trivial. Input is obtained via a microphone array over a socket. Output is published via a socket as well to a real-time virtual director represented by orchestration engine [3].

2.1. Audio capture

The core audio device for the system is a diamond array with four microphones (Figure 1) or a circular array with eight microphones. This can be handled with a VST host, a Data-Flow Controller (DFC) [4] or Audio Communication Engine (ACE) [5], which is a jitter adaptive, low delay audio streaming system, comprising several codec types. These programs act as servers, allowing processing modules requiring audio to connect over a socket and be supplied with audio at different sample rates.

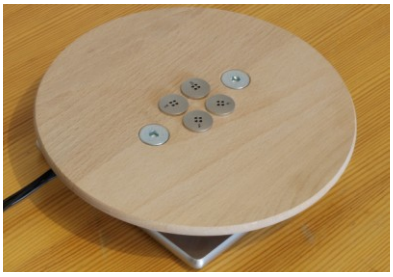

Figure 1: *Diamond array based on 4 omnidirectional microphones AKG C562CM.*

The socket solution (Figure 2) allows a very natural organisational interface. Usage of 10 ms packets results in keeping transmission latency within 12-20 ms.

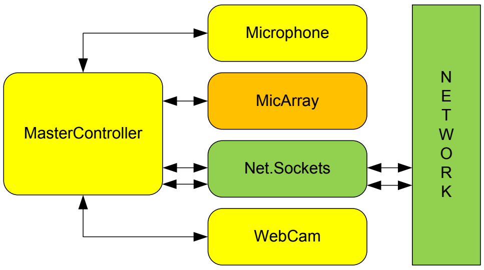

Figure 2: *Architecture of data acquisition server [4].*

2.2. Multi-framing data-flow processing

To combine the different feature extraction techniques, a data-flow architecture Tracter [6] is used. Data-flow is a well established signal processing technique that represents individual processing elements as vertices in a directed graph (Figure 3). The data is propagated through the graph using a “pull” mechanism, instigated by the sink. The request from the sink is propagated back through the network, with each element in turns requesting enough data from its inputs to perform the operation. Pull mechanisms lend themselves to systems that do not necessarily run in real-time. In this case, it allows the dataflow to be driven by the Weighted Finite State Transducer (WFST) decoder, which in turn is the most CPU-intensive task. Whilst it runs overall in real-time, it is not mandated to do so, as it would be by a push mechanism.

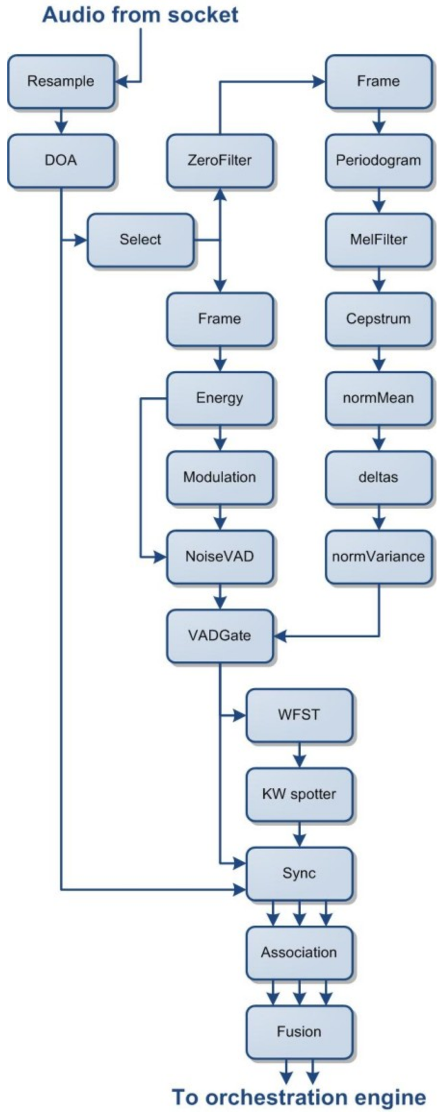

Figure 3: *Data-flow graph of the multichannel feature extraction, detection, keyword spotting, association and fusion.*

The data-flow graph operates in multi-framing mode with overlapped frames of 16 ms in step of 10 ms for ASR and frames of 32 ms in step of 16 ms for direction of arrival estimation (Figure 4). Audio packets from home entertainment server are retrieved each 10 ms and contain interleaved 4+ channel PCM audio in 16-bit sampled at 48 kHz or 16 kHz.

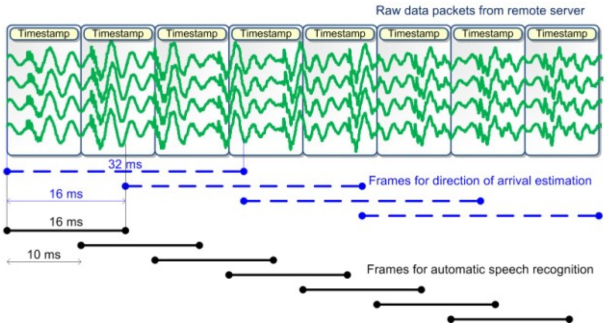
Raw data packets from remote server
Timestamp
Timestamp
Timestamp
Timestamp
Timestamp
Timestamp
Timestamp
Timestamp
32 ms
16 ms
Frames for direction of arrival estimation
16 ms
10 ms
Frames for automatic speech recognition

Figure 4: *Slicing frames for online processing.*

Each packet contains also unique 64-bit timestamp in microseconds for synchronisation between different remote modules. The results are published with new timestamps calculated by time countdown from the most recent timestamp. This allows to avoid time skew problem which can be observed due to unsynchronised clocking of different remote devices.

2.3. Direction of arrival estimation

Speaker localisation is performed by direction of arrival (DOA) plug-in (Figure 3). It can be effectively used with different types of microphone arrays. The algorithm is based on a generic sector-based activity measure (SAM) that relies only on the geometry of the microphone array [7] (as opposed to other techniques, such as [8], depending also on prior knowledge of the room dimensions). It is able to both detect and localise multiple sources. Fusion of directional clusters based on particle filter with time length of 80 ms and 15° jittering together with voice activity detection provides robust performance for detection and localisation of several concurrent speakers even in a reverberant environment.

Figure 5 shows the localised source positions for the setup with two fixed sound sources at angles -45° and +45° with respect to the microphone array. The array is placed at a distance *s = 20 cm* from the center between both speakers. The first speaker (+45°) is active between 0 < t < 20 s and 40 s < t < 60 s, whereas the second one (-45°) is active between 20 s < t < 60 s. There is a cross-talk situation during the last 20 s.

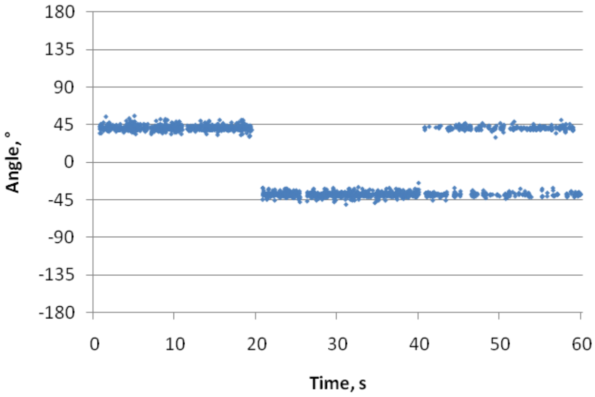

Figure 5: *Localised source positions (s=20cm, $\varphi_{L,R}$=±45°).*

Figure 6 shows the localised source positions for the setup with two fixed sound sources at angles -75° (first speaker) and +75° (second speaker) with respect to the microphone array. The array is placed at a distance *s = 80 cm*. The speaker activity times are as before.

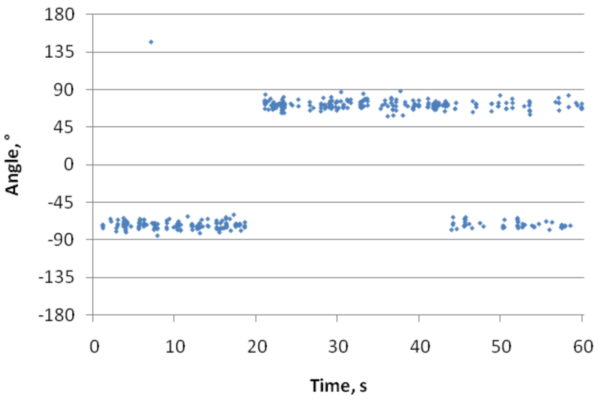

Figure 6: *Localised source positions (s=80cm, $\varphi_{L,R}$=±75°).*

The plots show that the increased reverberation reduces the localisation performance. However, the localisation still yields reliable results. The results (φ – angle, σ – standard deviation, L – left speaker, R – right speaker) for both cases are summarised in Table 1.

Table 1. *Statistics for localised source positions.*

Setup	φ_L	φ_R	σ_L	σ_R
s = 20 cm, $\varphi_{L,R}$=±45°	-38.9°	+41.5°	2.3°	2.4°
s = 80 cm, $\varphi_{L,R}$=±75°	-72.2°	+71.6°	4.1°	4.7°

It is worth mentioning that standard deviation does not increase in case of cross-talk, nevertheless it strongly depends on the reverberation level.

In addition, a probabilistic framework [9] can be used to determine the trajectories of multiple moving speakers in the short term while they speak. Instantaneous location estimates that are close in space and time are grouped into "short-term clusters" in a principled manner. Each short-term cluster determines the precise start and end times of an utterance, and a short-term spatial trajectory.

DOA plug-in is interchangeable with directional plug-in based on directional audio coding parameters [10], though the last one operates only at 44.1 and 48 kHz, while DOA plug-in can operate already at 16 kHz.

2.4. Feature acquisition

Audio is resampled to mono 16 kHz and pre-emphasised to flatten the spectral shape. A 256 point Discrete Fourier Transform (DFT) is performed in steps of 10 ms and squared to give the power spectrum. The resulting 129 unique bins are then decimated using a filter-bank of 23 overlapping triangular filters equally spaced on the mel-scale to model human auditory system. A logarithm and DFT then yield the mel-cepstrum, which is truncated, retaining the lower 13 dimensions. This truncation retains spectral shape and discards excitation frequency. Next, Cepstral Mean Normalisation (CMN) is performed by subtracting from each cepstral vector the mean of the vectors of the preceding (approximately) half second. This has the effect of removing convolutional channel effects. Finally, the 13 normalised cepstral coefficients are augmented by first and second order derivatives, corresponding to their velocity and acceleration. This gives 39 dimensional vectors.

2.5. Voice activity detection

Voice activity detection (VAD) covers both verbal and paralinguistic activity and is implemented in Tracter as a gate. Downstream from the gate, the decoder is unaware that VAD is happening; it just receives segmented data as if it were reading from a sequence of pre-segmented utterances (files). Upstream from the gate, however, the data is actually one continuous stream.

The gate segments the input stream based upon boolean speech / non-speech information from a VAD algorithm based on silence models [11] for WFST. A simple energy based VAD is interchangeable with an MLP based VAD plug-in trained such that typical ASR features on the input layer appear as speech and silence outputs at the output layer.

2.6. WFST decoder and keyword spotter

The ASR component is represented by WFST based token passing decoder known as Juicer [12]. The output from Juicer is used to perform spotting of proper names and keywords. This is still an active research topic and has a high level of complexity in the context of a living room environment. In real life conditions the system has to cope with a quite high level of acoustic background "noise" and speaker-independency.

Juicer is architecturally a Tracter sink. This means that any Tracter graph can be used seamlessly for feature acquisition. Juicer can operate directly on high order language models in real-time. We use 3-gram, although higher order is possible.

The real-time system uses the language models developed for the NIST RT evaluations [13]. These are typically 50K word N-gram models. Generally speaking, although the final system will run on a 32 bit system, i.e., in under 4GB of memory, the WFSTs must be composed on a 64-bit system. Table 2 shows positioning of the WFST decoder Juicer in real time on the RT07 test set. For comparison, we also give the (multi-pass) AMI result on the more recent RT09 evaluation [14], which can be taken as a lower bound on the achievable error rate for Individual Headset Microphone (IHM) and Multiple Distant Microphone (MDM).

Table 2. *Word error rate of Juicer in RT07 and RT09 evaluations at NIST.*

Evaluation	*IHM*	*MDM*
RT07, real-time	37%	41%
RT09, multi-pass	23.5%	33.2%

Spotting of proper names and keywords is performed based on the list of participant proper names and keywords relevant to the enabled scenario. This list is currently predefined, nevertheless can be provided in the future by virtual director.

2.7. Low delay data association and fusion

Due to the real-time requirements, association and fusion of directional information with voice activity cannot be postponed till voice activity is over. The fused events have to be published to virtual director within few hundred ms to keep the feeling of virtual presence. The actual low delay association scheme is depicted on Figure 7.

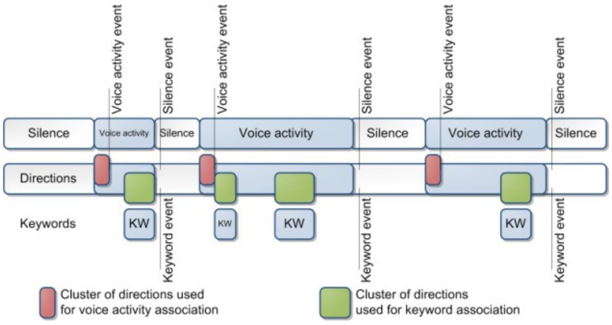

Figure 7: *Low delay association and fusion.*

In case the voice activity confirmation window is set to 150 ms, only 150 ms of directional information can be used at the time of voice activity event transmission to the orchestration engine. In reality, the directional particle filter reduces this window to 86 ms, because it needs 32 ms at each end to initialise the cluster. The results (τ – latency, υ – maximum number of confirmed elements in directional cluster) for different voice activity confirmation length are summarised in Table 3.

Table 3. *Latency and directional cluster size for voice activity low delay association.*

Confirmation length	τ	υ
150 ms	*150 ms*	4
200 ms	*200 ms*	7
250 ms	*250 ms*	9
300 ms	*300 ms*	12

The delay of silence event equals to silence confirmation window and can vary from 50 ms to 300 ms. Association of silence is performed with directional cluster used for voice activity event. Successfully associated events are assigned with high confidence, while other events are assigned with low confidence (e.g., in case of no confirmed elements in corresponding directional cluster).

Proper names and keywords are spotted at silence confirmation time for the preceding voice activity segment, therefore time window of each proper name / keyword can be used according to the start and stop timestamps.

2.8. Output

The XML encoded results representing time-stamped events are published via XML-RPC protocol to virtual director represented by orchestration engine with interaction ontology, which makes decisions on the selection and transformation of audiovisual content.

One of employed scenarios is family game, played by spatially separated friends and families.

3. Conclusions

We have developed a real-time low delay system for hands free audio analysis for a living room environment. It comprises detection and localisation of the verbal and paralinguistic events, which can augment the behaviour of virtual director and improve the overall experience of interactions between spatially separated families and friends. Whilst none of the system components are new in themselves, we have shown how they can be joined together to form a working system. For each component, measured or expected performance has been analysed to give an overall feel for what might be expected from state of the art components arranged in such a way. The achieved results allow us in the future to enlarge the system towards audiovisual processing by inclusion of multiple face tracking, spatial association and multimodal fusion.

4. Acknowledgements

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. Numerous people have contributed, with and without their knowledge, to the system described above. Juicer and its associated tools were originally written by Darren Moore and the direction of arrival software by Iain McCowan. We are grateful to Fraunhofer IIS for provision of the real life dataset and to Sebastian Vogel for resolving ACE compatibility issues.

5. References

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

[2] Garner, P.N., Dines, J. "Tracter: A Lightweight Dataflow Framework", Proceedings of Interspeech, Makuhari, Japan, 2010.

[3] Williams, D., Ursu, M.F., Cesar, P., Bergstrom, K., Kegel, I., Meenowa, J., "An emergent role for TV in social communication", Proceedings of the 7th European Conference on European Interactive Television EuroITV 2009, Leuven, Belgium, 2009.

[4] Korchagin, D., "Multimodal Data Flow Controller", Idiap-Com-01-2009, Martigny, Switzerland, 2009.

[5] Issing, J., Reuschl, S., Farber, N., German, R., "RTCP based Bit-Rate Adaptation for AAC Audio Communication", in Proc. NEM Summit 2009, Saint Malo, France, 2009.

[6] Open source data flow framework "Tracter", http://juicer.amiproject.org/tracter.

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

[8] Duraiswami, R., Zotkin, D. and Davis, L.S., "Active Speech Source Localization by a Dual Coarse-to-Fine Search", in IEEE Proc. ICASSP, 2001.

[9] Lathoud, G. and Odobez, J.-M., "Short-Term Spatio-Temporal Clustering Applied to Multiple Moving Speakers", in IEEE Transactions on Audio, Speech and Language Processing, 2007.

[10] Thiergart, O., Schultz-Amling, R., Del Galdo, G., Mahne, D. and Kuech, F., "Localization of sound sources in reverberant environments based on directional audio coding parameters", in 127th AES Convention, New York, USA, 2009.

[11] Garner, P. N., "Silence models in weighted finite-state transducers", in Proceedings of Interspeech, Brisbane, Australia, 2008.

[12] Garner, P.N., Dines, J., Hain, T., Hannani, A.E., Karafiat, M., Korchagin, D., Lincoln, M., Wan, V., Zhang, L., "Real-Time ASR from Meetings", Proceedings of Interspeech, pp. 2119-2122, Brighton, UK, 2009.

[13] Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Leeuwen, D., Lincoln, M. and Wan, V., "The 2007 AMI(DA) system for meeting transcription", in Multimodal Technologies for Perception of Humans, ser. Lecture Notes in Computer Science. Springer-Verlag, 2008, vol. 4625, pp. 414-428, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers.

[14] Hain, T., Burget, L., Dines, J., Garner, P.N., Hannani, A.E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., "The AMIDA 2009 Meeting Transcription System", Proceedings of Interspeech, Makuhari, Japan, 2010.