AN ONLINE SPEAKER TRACKING SYSTEM FOR AMBIENT INTELLIGENCE ENVIRONMENTS

Maider Zamalloa, Mikel Penagarikano , Luis Javier Rodríguez-Fuentes, Germán Bordel
GTTS, Department of Electricity and Electronics, University of the Basque Country, Leioa, Spain
maider.zamalloa@ehu.es, mikel.penagarikano@ehu.es, luisjavier.rodriguez@ehu.es, german.bordel@ehu.es

Juan Pedro Uribe
Ikerlan - Technological Research Centre, Mondragón, Spain
JPUribe@ikerlan.es

Keywords: Speaker Tracking, Ambient Intelligence, AMI Corpus

Abstract: Ambient intelligence is an interdisciplinary paradigm which envisages smart spaces that provide services and adapt transparently to the user. As the most natural interface for human interaction, speech can be exploited for adaptation purposes in such scenarios. Low latency is required, since adaptation must be continuous. Most speaker tracking approaches found in the literature work offline, fully processing pre-recorded audio files by a two-stage procedure: (1) performing acoustic segmentation and (2) assigning each segment a speaker label. In this work a real-time low-latency speaker tracking system is presented, which deals with continuous audio streams. Experimental results are reported on the AMI Corpus of meeting conversations, revealing the effectiveness of the proposed approach when compared to an offline speaker tracking system developed for reference.

1 INTRODUCTION

Ambient Intelligence (AmI) is an interdisciplinary applied research field, aiming to create smart spaces which provide services featuring user and context adaptation capabilities (ISTAG, 2001) (Cook, 2009). It was originally devised as *Ubiquitous Computing* in (Weiser, 1991) where it was suggested the interaction of consumer electronics, telecommunications and computing devices to support people carrying out everyday life activities in a natural way. In such environment, daily objects feature computing and telecommunication capabilities. *Transparency* is critical, so natural and intelligent interfaces are needed for human-computer interaction (Abowd, 2005). Speech is a natural interface for human interaction and the most suitable means to support user interaction and adaptation. Speech streams can be exploited to extract user related information such as location, identity, etc. But in such environments, user adaptation must be continuous, and low-latency online processing is needed.

Speaker diarization and speaker tracking are well known tasks which aim to answer the question *Who spokes when?*. Speaker tracking aims to detect segments corresponding to a known set of target speakers (Martin, 2001). Speaker diarization consists of detecting speaker turns without any prior knowledge about the target speakers (Tranter, 2006) (Meignier, 2006). Speaker diarization and tracking primary applications domains assume that audio recordings are fully available before processing. Common approaches to these tasks consist of two uncoupled steps: (1) audio segmentation and (2) speaker detection. In speaker diarization, segments hypothetically uttered by the same speaker are clustered together. In speaker tracking, however, once the audio stream is segmented, speaker detection is carried out through classical speaker recognition techniques (Moraru, 2005) (Istrate, 2005) (Bonastre, 2000). In any case, these methodologies are not suitable for low-latency online speaker detection.

Few works related to real-time speaker segmentation and tracking can be found in the literature (Wu, 2003) (Lu, 2005) (Liu, 2005). Most of the speaker segmentation approaches are based on metrics measuring spectral changes, such as Bayesian Information Criterion (Chen, 1998) and Generalized Likelihood Ratio (Bonastre, 2000) (Liu, 2005). These procedures are robust but computationally expensive since two or three Gaussian models must be estimated for scoring each possible change point in each analysis window, and there can be between 100 and 1000 analysis windows per second. In (Wu, 2003), a real-time model-based speaker change detection system is proposed, where a Universal Background Model (UBM) is taken as reference to classify speech segments, and a distance between two adjacent windows is computed which accounts for the spectral change. In (Lu, 2005), an unsupervised speaker segmentation and tracking algorithm is presented. Once the speaker change boundaries are determined, each segment is scored with a set of incremental quasi Gaussian Mixture Models corresponding to unknown target speakers. In (Liu, 2005), an online speaker adaptation methodology is applied for real-time speaker tracking, with unknown target speakers. This approach combines a phonotactic speaker change detection module with an online speaker clustering algorithm. Speaker adaptation is based on feature transformation. The transformation matrix is incrementally adapted as labeled segments become available.

In this paper, a real-time low-latency online speaker tracking approach is presented, designed for an AmI scenario (for example, an intelligent home environment), where the system continuously tracks known speakers. The expected number of target speakers is low (i.e. the members of a family). This scenario requires taking almost instantaneous (low latency) speaker tracking decisions. A very simple speaker tracking algorithm is proposed, where audio segmentation and speaker detection are jointly accomplished by defining and processing fixed-length audio segments and scoring each of them to decide whether it belongs to a target speaker or to an impostor. Audio segments are scored by means of acoustic models (corresponding to target speakers) estimated via Maximum a Posteriori (MAP) adaptation of a UBM (Reynolds, 2000). The MAP-UBM methodology yields good speaker recognition performance and allows for a fast scoring technique which speeds up the score computation. Finally, detection scores are calibrated (i.e linearly mapped to likelihood ratios) and then, based on the scores obtained for a development corpus, an optimal application-dependent decision threshold (that minimizes the expected error cost) is established (Brummer, 2006). The performance of the proposed approach is compared to that of an offline system developed for reference, which follows the classical two-stage approach: audio segmentation is done over the whole input stream, and MAP-UBM speaker detection is performed on the resulting segments. Speaker tracking experiments applying both systems were carried out on the AMI Corpus (Carletta, 2007), which contains human conversations in the context of smart meeting rooms, close to the AmI scenario described above.

The rest of the paper is organized as follows. In section 2 the main features of the speaker tracking systems are described, including the acoustic front-end, the audio segmentation (required for the offline reference system) and the speaker detection and calibration stages. Section 3 gives details about the experimental corpus and the UBM estimation. Speaker tracking results using the proposed and the reference systems are presented in section 4. Finally, conclusions and guidelines for future work are given in Section 5.

2 SPEAKER TRACKING SYSTEMS

2.1 Acoustic Front-End

In this work, 16 kHz audio streams are analyzed in frames of 20 milliseconds, at intervals of 10 milliseconds. A Hamming window is applied and a 512-point FFT computed. The FFT amplitudes are then averaged in 24 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform is finally applied to the logarithm of the filter amplitudes, obtaining 12 Mel-Frequency Cepstral Coefficients (MFCC). To increase robustness against channel distortion, Cepstral Mean Normalization (CMN) is applied. When the audio stream is processed on-the-fly, a dynamic CMN approach is applied, the cepstral mean being updated at each time t as follows:

$$\mu_t = \alpha C(t) + (1-\alpha)\mu_{t-1} \qquad (1)$$

where α is a time constant, C(t) is the vector of cepstral coefficients at time t and μ_{t-1} is the dynamic cepstral mean at time t-1. After CMN, the first derivatives of the MFCC are also computed, yielding a 24-dimensional feature vector.

2.2 Audio Segmentation

Audio segmentation, also known as acoustic change detection, is required by most speaker tracking systems as a previous step to the detection of target speakers. A simple algorithm is applied in this work, which segments the audio signal in a fully unsupervised way, by locating the most likely change points from a purely acoustic point of view. The algorithm considers a sliding window W of N acoustic vectors and computes the likelihood of change at the center of that window, then moves the window K vectors ahead and repeats the process until the end of the vector sequence. To compute the likelihood of change, each window is divided in two halves, W_{left} and W_{right}, then a Gaussian distribution (with diagonal covariance matrix) is estimated for each half and finally the cross-likelihood ratio is computed and stored as likelihood of change. This yields a sequence of cross-likelihood ratios which must be post-processed to get the hypothesized segment boundaries. This involves applying a threshold τ and forcing a minimum segment size δ. In practice, a boundary t is validated when its cross-likelihood ratio exceeds τ and there is no candidate boundary with greater ratio in the interval $[t-\delta,t+\delta]$ (see (Rodriguez, 2007) for details).

2.3 Speaker Detection

The real-time speaker tracking system proposed in this work computes a detection score per target speaker and outputs a speaker identification decision at fixed-length intervals. That length has been empirically set to one second, which provides relatively good time resolution and spectral richness, and a reasonably small latency for most online speaker tracking scenarios. The offline system developed for reference does the same computation, but using the segments produced by the algorithm described in Section 2.2. Regardless the way audio segments are obtained, they are scored with the same set of MAP-UBM target speaker models (Reynolds, 2000).

In the adaptation of a speaker model from the UBM, only *non-overlapped* training segments (i.e. those segments containing only speech from that speaker, according to the time references of manual annotations) are used. This way, component densities related to the acoustic classes strongly observed in training data will change, whereas component densities that correspond to weaker or missing acoustic units (such as silence or impostors) will remain un-adapted. Therefore, it is assumed that the resulting MAP-UBM system should be able to detect speech from target speakers and reject silence, noise and speech from impostors.

Given the acoustic model λ_s for the target speaker s and λ_{UBM} for the UBM, the detection score $\Delta_s(X)$ is computed as follows:

$$\Delta_s(X)= L(X|\lambda_s) - L(X|\lambda_{UBM}) \qquad (2)$$

where $L(X|\lambda)$ is the log-likelihood of X given λ. Once the detection scores are computed for all the target speakers, a unique identification decision is made per segment: X is marked as coming from the most likely target speaker $s^* = \arg\max_{s\epsilon[1,S]}\{\Delta_s(X)\}$, if $\Delta_{s^*}(X) > \theta$. Otherwise X is marked as coming from an impostor. The decision threshold θ can be heuristically established to optimize the discrimination. Note that, for any given segment X, there could actually be two or more speakers speaking at the same time. However, the detection approach described above cannot inform of speaker overlaps, because only the most likely speaker can be detected.

2.4 Calibration of scores

Calibration maps detection scores $\{\Delta_s \mid s \in [1,S]\}$ to likelihood ratios $\{C(\Delta_s) \mid s \in [1,S]\}$ without any specific application in mind. The scaling parameters are computed over a development corpus by maximizing *Mutual Information*, which is equivalent to minimizing the so called C_{LLR} (a metric defined in (Brummer, 2006)), which integrates the expected cost over a wide range of operation points (representing specific applications) in the Detection Error Tradeoff (DET) curve (Martin, 1997). The final decision is taken by applying the minimum expected cost Bayes decision threshold to calibrated scores $C(\Delta)$. The target speaker is accepted only if the following inequality holds:

$$C(\Delta)\geq \ln\left(\frac{C_{fa}(1-P_{target})}{C_{miss}P_{target}}\right) \qquad (3)$$

where C_{miss} and C_{fa} are miss and false-acceptance error costs, and P_{target} the prior probability of target speakers. Scores are calibrated by means of the *FoCal toolkit*, applying a linear mapping strategy (see http://www.dsp.sun.ac.za/~nbrummer/focal/).

3 EXPERIMENTAL SET-UP

3.1 The AMI Corpus

Experiments are carried out over the AMI Corpus of meeting conversations, available as a public resource (see http://corpus.amiproject.org/). The AMI Corpus is a multimodal dataset concerned with real-time human interaction in the context of smart meeting rooms. Data, collected in three instrumented meeting rooms, include a range of synchronized audio and video recordings. Meetings contain speech in English, mostly from non native speakers.

In this work, the development and evaluation of speaker tracking systems is based on a subset of the AMI Corpus, the Edinburgh scenario meetings, including 15 sessions: ES2002-ES2016, with four meetings per session, each meeting being half an hour long on average. Training data are taken from meetings recorded at the three AMI sites. The audio stream is obtained by mixing the signals from the headset microphones of the participating speakers. Three of the four speakers participating in each session are taken as target speakers, the remaining one being assigned the role of impostor. Careful impostor selection –not random– is made to account for gender unbalanced sessions. In sessions containing just one female speaker, the impostor is forced to be male (and vice versa), in order to avoid that gender favors impostor discrimination.

In order to assess the speaker tracking performance in realistic conditions, two independent subsets are defined, consisting of different sessions (and therefore different speakers), for development and evaluation purposes, respectively. The development set, consisting of 8 sessions (32 meetings), is used to tune the configuration parameters of the speaker tracking systems. The evaluation set, including the remaining 7 sessions (28 meetings), is used only to evaluate the performance of the previously tuned speaker tracking systems.

Both the development and evaluation subsets are further divided into train and test datasets. Two meetings per session are randomly selected for training speaker models, and the remaining two are left for testing purposes. Time references are based on manual annotations provided in the AMI Corpus.

3.2 UBM estimation

Two speaker detection systems have been developed based on the MAP-UBM approach. They only differ in the data used to estimate the UBM: UBM-g uses 15 gender-balanced AMI meetings from all sites except Edinburgh (so, a kind of room mismatch may be expected), whereas UBM-t uses only speech from target speakers in training meetings. UBM-g is estimated once and can be applied to whatever evaluation data and target speakers, whereas UBM-t must be estimated specifically for each set of target speakers.

3.3 Performance measures

The performance of speaker tracking systems is commonly analyzed by means of *Detection Error Tradeoff* (DET) plots (Martin, 1997). Performance is measured in terms of time that is correctly or incorrectly classified as belonging to a target. Therefore, miss and false alarm rates are computed as a function of time (Martin, 2001) and not as a function of trial number, like in speaker detection experiments.

DET performance can be summarized in a single figure by means of the *Equal Error Rate* (EER), the point of the DET curve at which miss and false alarm rates are equal. Obviously, the lower the EER, the higher the accuracy of a speaker tracking system.

Another way to summarize in a single figure the performance of a speaker tracking system is the so called *F-measure*, defined as follows:

$$F=\frac{2.0*PRC*RCL}{PRC+RCL} \qquad (4)$$

where precision (PRC) and recall (RCL) are related to false alarm and miss rates respectively. PRC measures the correctly detected target time from the total target time detected. RCL computes the correctly detected target time from the actual target time. The F-measure ranges from 0 to 1, with higher values indicating better performance. Collar periods of 250 milliseconds at the end of speaker turns are ignored for scoring purposes. Thus, speaker turns of less than 0.5 seconds are not scored.

4 EXPERIMENTAL RESULTS

Figures 1 and 2 show the performance of the online and offline speaker tracking systems, using UBM-g and UBM-t as background models, for the development and evaluation sets, respectively. Since the speaker detection strategy followed in this work cannot detect speaker overlaps, all the segments containing speech from two or more speakers are removed when scoring test meetings.

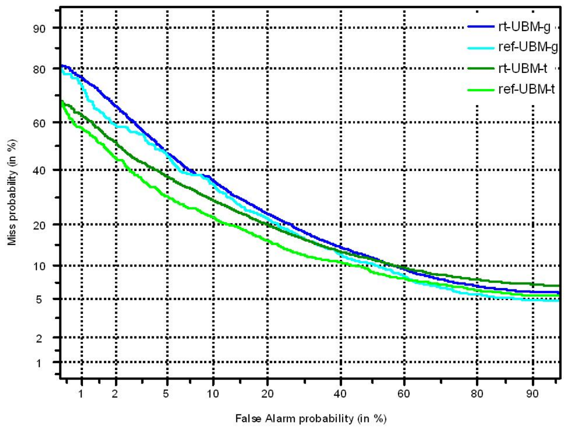

Figure 1: DET performance of speaker tracking systems on the development set defined on the AMI corpus.

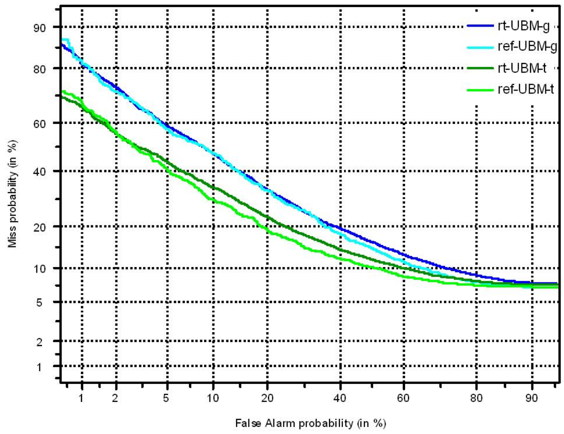

Figure 2: DET performance of speaker tracking systems on the evaluation set defined on the AMI corpus.

As expected, the classical offline system outperforms the proposed low-latency online system, but the performance of the latter is quite good. Taking the performance of the offline system as reference, in speaker tracking experiments over the evaluation set, the EER increases from 25.80 to 26.20 (1.55% relative degradation) when using UBM-g, and from 19.07 to 21.14 (10.85% relative degradation) when using UBM-t. On the other hand, UBM-t systems outperform UBM-g systems, maybe due to the aforementioned room mismatch in UBM-g and the limited amount of training data.

In addition, performance degradation from development to evaluation is small (from around 20% to 21% EER) in MAP-UBM-t, which means that system configuration (based on the development set) was also suitable for the evaluation set. The MAP-UBM-g system suffers a bigger degradation from development to evaluation. It seems that estimating the UBM from unknown speakers in mismatched conditions (different rooms) degrades acoustic coverage and reduces the robustness of system configuration with regard to using a room-specific UBM estimated from target speakers. The UBM-t system might be getting advantage not only from matching the room, but also from the consistency between the speakers in the UBM and the target speakers. In fact, 100% of the target speakers appearing in the test corpus contribute data to the UBM-t, increasing the consistency of speaker models estimated through MAP adaptation (because a perfect match exists between the adaptation data corresponding to any target speaker and some of the component densities of the UBM).

Table 1 shows precision (PRC), recall (RCL) and F-measure performance of the speaker tracking systems, for both the calibrated and uncalibrated speaker detection scores. These results correspond to the operation point (threshold) considered optimal in the DET curve. The threshold used for calibrated scores is based on application-dependent costs and target priors, adjusted on the development corpus. For uncalibrated scores, the threshold is fixed to zero, i.e. a target speaker is detected if the likelihood of the null hypothesis is higher than that of the alternative hypothesis.

Table 1: Precision (PRC), Recall (RCL) and F-measure performance of the real-time (rt) and reference (ref) speaker tracking systems, using UBM-g and UBM-t, on the development (Dev) and evaluation (Eval) sets.

		Uncalibrated			Calibrated		
		PRC	RCL	F	PRC	RCL	F
Dev	rt-UBM-g	0.66	0.92	0.77	0.81	0.8	0.81
	ref-UBM-g	0.67	0.93	0.78	0.82	0.82	0.82
	rt-UBM-t	0.67	0.91	0.77	0.82	0.83	0.82
	ref-UBM-t	0.69	0.92	0.79	0.84	0.86	0.85
Eval	rt-UBM-g	0.69	0.92	0.78	0.78	0.85	0.8
	ref-UBM-g	0.69	0.93	0.79	0.78	0.84	0.81
	rt-UBM-t	0.71	0.91	0.8	0.81	0.85	0.83
	ref-UBM-t	0.72	0.92	0.81	0.81	0.87	0.84

Results in Table 1 demonstrate the usefulness of the calibration stage, which leads to better performance in all cases. Finally, note that the real-time (online, low-latency) system provides only slightly worse performance than the reference (offline) system: 1.7% average relative degradation in F-measure. Though speaker tracking actually takes advantage from an offline acoustic segmentation of the audio stream, depending on the scenario and the required latency, offline audio segmentation would not be feasible. In such a situation, the proposed approach provides real-time low-latency online speaker tracking at the cost of little performance degradation.

5 CONCLUSIONS AND FUTURE WORK

In this paper, an online speaker tracking system, designed for an Ambient Intelligence scenario, is presented an evaluated. The system processes continuous audio streams and outputs a speaker identification decision for fixed-length (one second) segments. Speaker detection is done by means of a MAP-UBM speaker verification backend. A calibration stage is applied which linearly maps detection scores to likelihood ratios. Calibration parameters are estimated beforehand based on development data, yielding significant performance improvements without increasing the computational cost, which is crucial for a real-time low-latency system. An alternative speaker tracking system, based on an offline segmentation of the audio stream has been developed and evaluated for reference.

Experiments have been carried out on a subset of the AMI Corpus of meeting conversations. Results demonstrate that better results can be attained when the UBM is estimated from data matching test conditions (same room, same speakers), instead of using general but unrelated data. The calibration stage provides performance improvements in all cases. Finally, offline segmentation of audio streams actually improves speaker tracking performance with regard to using fixed-length segments. However, depending on the scenario and the required latency, offline audio segmentation would not be feasible. The proposed system provides real-time low-latency online speaker tracking with little performance degradation.

Current work involves increasing the robustness of detection scores (and decisions) by using information from past segments. Future work includes using detection scores in a speaker verification framework (thus allowing the detection of multiple speakers), and making a smart use of all the available data through new UBM estimation strategies.

ACKNOWLEDGEMENTS

This work has been partially funded by the Government of the Basque Country, under program SAIOTEK, project S-PE07IK03; and the Spanish MICINN, under Plan Nacional de I+D+i, project TIN2009-07446.

REFERENCES

Abowd, G.D., Mynatt, E.D., "Designing for the Human Experience in Smart Environments" in D.J. Cook and S.K. Das, Editors, Smart Environments: Technology, Protocols, and Applications, Wiley, 153-174, 2005.

Bonastre, J.F., Delacourt, P., Fredouille, C., Merlin, T. and Wellekens, C., "A Speaker Tracking System based on Speaker Turn Detection for NIST Evaluation", in Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, 2000.

Brummer, N. and Preez, J., "Application Independent Evaluation of Speaker Detection", Computer Speech and Language, 20:230-275, 2006.

Carletta, J., "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus". Language Resources and Evaluation Journal , 41(2): 181-190, 2007.

Chen, S.C. and Gopalakrishnan, P.S., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

Cook, D.J., Augusto, J.C., Jakkula, V.R., "Ambient Intelligence: Technologies, Applications, and Opportunities", Pervasive and Mobile Computing, 5(4): 277-298, 2009.

ISTAG, "Scenarios for Ambient Intelligence in 2010". European Commission Report, 2001.

Istrate, D., Scheffer, N., Fredouille, C. and Bonastre, J.F., "Broadcast News Speaker Tracking for ESTER 2005 Campaign", in Proceedings of the International Conference on Speech and Language Processing, Lisboa, 2005.

Liu, D., Kiecza, D., Srivastava, A. and Kubala, F., "Online Speaker Adaptation and Tracking for Real-time Speech Recognition", in Proceedings of the International Conference on Speech and Language Processing, Lisboa, 2005.

Lu, L. and Zhang H.J., “Unsupervised Speaker Segmentation and Tracking in Real-time Audio Content Analysis”, Multimedia Systems,10:332-343, 2005.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., "The DET curve in assessment of detection task performance", in Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), Vol. 4, pp. 1895-1898, 1997.

Martin, A.F. and Przybocki, M.A., "Speaker Recognition in a Multi-Speaker Environment", in Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), Denmark, 2001.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F. and Besacier, L., "Step-by-step and Integrated Approaches in Broadcast News Speaker Diarization", Computer Speech and Language, 20:303-330, 2006.

Moraru, D., Ben, M., Gravier, G., "Experiments on Speaker Tracking and Segmentation in Radio Broadcast News", in Proceedings of the International Conference on Speech and Language Processing, Lisboa, 2005.

Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., "Speaker Verification Using Adapted Gaussian Mixture Models". Digital Signal Processing, 10:19-41, 2000.

Rodríguez, L.J., Peñagarikano, M. and Bordel, G., "A Simple But Effective Approach to Speaker Tracking in Broadcast News", Pattern Recognition and Image Analysis*,* LNCS 4478: 48-55, Springer-Verlag, 2007.

Tranter, S.E. and Reynolds, D.A., "An Overview of Automatic Speaker Diarization Systems", IEEE Transactions on Audio, Speech and Language Processing, 14(5): 1557-1565, 2006.

Weiser, M., "The Computer for the Twenty-First Century", Scientific American, 94-104, 1991.

Wu, T.Y., Lu, L., Chen, K. and Zhang, H.J., "UBM-based Real-time Speaker Segmentation for Broadcasting News", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, China, 2003.