VOICE SEPARATION OF OVERLAPPING SPEECH USING TRACKING TECHNIQUES AND THE GATING PROCESS

Ilyas Potamitis, Panos Zervas, Nikos Fakotakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
potamitis@wcl.ee.upatras.gr

ABSTRACT

This paper investigates the use of tracking techniques successfully applied to aircraft tracking and navigation to segment possibly overlapping speech of multiple static speakers in an enclosure. The tracking technique applied, namely the probabilistic data association (PDA) in conjunction with the interacting multiple model (IMM) estimator directly accounts for measurement origin uncertainty, i.e., which direction of arrival (DOA) measurement comes from which speaker and rejects spurious DOAs. The estimated DOAs are utilized by a single microphone array to provide separation through its directional receptive field. Based on the prediction of the IMM filter that constructs permissible DOA regions for each speaker (gates), we elaborate on the concept and application of the so called 'gating process' that can be utilized in the initialization and termination of speech tracks, thus serving as a voice activity detector (VAD). The effectiveness of the approach is illustrated by extensive simulation study on tracking and separating three static speakers having a conversation with partially overlapping speech and long pauses.

1. INTRODUCTION

Microphone arrays are electronically steerable, angle-of-arrival filters, designed to constrain their receptive field to a desired direction (i.e., beamforming process) [1], [2]. Their ability to provide spatially selective speech acquisition makes them good candidates for a wide field of applications including teleconferencing, multimedia conferencing [3-4], speech recognition with regard to distant talkers [5] and automatic camera steering [4], [6]. The majority of reported research deals with beamforming a single speaker [3]-[6]. The multi-speaker case is either treated as a single beam switching to the stronger speaker or limited to multi-speaker scenarios where the DOAs of the speakers' wavefronts are assumed known. The main difficulty of the multiple speakers' case is that the beamforming process becomes severely complicated due to the ambiguity of the origin of DOA measurements. The direct application of DOA-estimation techniques over consecutive frames does not yield tracking of each individual speaker, as beamforming on an extended basis would require. Moreover, multimedia, teleconference and 'smart' home applications of the future will have to deal with the complex interaction and speech overlap of participants as experienced in a normal, vivid conversation. In this work we will present a novel, general framework that can deal with the latter case.

In the context of speaker beamforming, a track is a trajectory of angles constructed by observations that have been associated with the same speaker. In order to be able to cover the unpredictable movement of the speaker over time, the proposed state inference scheme, the IMM [7], handles the uncertainty of the speaker's motion by incorporating multiple motion models in the tracking process. The workhorse of the estimation process is a bank of sequential Kalman-based tracking algorithms each incorporating a model for speakers' motion into the procedure of recursively deriving DOA estimates.

This work builds upon [8] where tracking techniques were applied to moving speakers. Although the latter case was more difficult it required that the speakers were active during the whole tracking period. In this work we extend the previous approach by making use of the prediction of the Kalman filters to initiate and terminate tracks of partially overlapping speakers that can have long silence parts in their utterances.

A data association technique is also incorporated into the state inference scheme to efficiently reject clutter measurements and to unambiguously associate the angle observations to speakers. The latter allows each receptive beam to track and lock on an extended basis to the same speaker using a single array and, therefore, to achieve a degree of separation of voices and reduction of reverberation (due to the spatial selectivity of the receptive field).

2. DOA ESTIMATION OF WIDEBAND SOURCES

The beamforming process applied on a sound source is usually implemented in two steps: a) the DOAs are derived from the impinging waveforms, and, b) a receptive beam is formed using an optimality criterion. There is a rich literature on DOA estimation techniques with different sensitivity to background noise and reverberation [1]-[2]. In general, the quality of the DOA measurements is affected by the following three factors:

a) The spectral content of the speech segment used to derive the DOA. An utterance is composed of a succession of high energy – low energy segments interspersed with silence parts. The low energy speech and silence parts are sensitive to background noise and are probable to return erroneous DOA measurements.

b) The reverberation level of the room can give rise to spurious measurements due to reflections on the walls and objects of the enclosure.
c) On the relative positioning of the array with respect to the talkers, the number of simultaneous sources present in the receptive field and their relative positioning.

This paper aims at incorporating kinematical models in the application of DOA estimation methods of speech signals in order to a) reject spurious angle measurements, b) reduce the variability of angle measurements and, c) associate angles to speakers and, therefore, allow voice separation and the association of consecutive speech segments of the same speaker. The proposed speech separation framework does not rely on a specific DOA estimation or beamforming algorithm. However the experimental results presented are computed using wideband MUSIC for DOA estimation of the moving sources and minimum variance for the beamforming process.

3. TRACKING PROCEDURE

An active speaker's trajectory can be subdivided into distinct segments, each corresponding to a different behavioral mode of movement. We use angle and angle rate as the state of a speaker's motion, and a discrete regime variable which describes the distinct segments of motion. All speakers have the same set of modes (each mode corresponds to a Kalman filter or an extended Kalman filter). Each filter corresponds to a behavior mode that undergoes jumps from model *i* to model *j,* according to a Markov transition matrix. The IMM algorithm merges the state estimates produced by each model at the beginning of each cycle. The problem of IMM state estimation in the context of our work is to infer the kinematical state based on noisy DOA measurements. One cycle of tracking at the IMMs is as follows [7], [9].

Step 0: An initial estimate of the speaker's angular location is provided by the DOA technique applied and an initial clustering of the angle values. Each of the speakers' state equation describing their movement and the observation equation is a linear function of the state. IMM consists of multiple (say *j*) models. We assume that the speakers follow each model at time *k* with probability μ_j. In this work the Newtonian source motion and observational equations become:

$$\mathbf{s}_j(k)=\mathbf{F}\mathbf{s}_j(k-1)+\mathbf{v}_j(k) \quad (1)$$

$$\mathbf{y}_j(k)=\mathbf{H}\mathbf{s}_j(k)+\mathbf{w}_j(k) \quad (2)$$

$\mathbf{v}_j(k)\sim N(\mathbf{0}, \mathbf{Q}_j)$ is the Gaussian zero mean process noise vector having covariance matrix $\mathbf{Q}_j$. $\mathbf{v}_j(k)$ models accelerations experienced by a moving source. $\mathbf{w}_j(k)\sim N(\mathbf{0}, \mathbf{R}_j)$ is the measurement noise with covariance $\mathbf{R}_j \sim 1/\sin^2\theta$.

Step 1 (Interaction): The previous outputs from each filter are combined through the previous model probabilities $\mu_j(k\text{-}1)$ to produce the mixed input $\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1)$ for the current cycle. That is

$$\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1)=\sum_i \mu_j(k\text{-}1)\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1) \quad (3)$$

$$\mathbf{P}_j(k\text{-}|k\text{-}1)=\sum_i \mu_j(k\text{-}1)[\mathbf{P}_j(k\text{-}1|k\text{-}1)+[\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1)-\hat{\mathbf{s}}(k\text{-}1|k\text{-}1)][\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1)-\hat{\mathbf{s}}(k\text{-}1|k\text{-}1)]^T] \quad (4)$$

Step 2: (Kalman Update and Prediction). Subsequently each filter functions as a simple Kalman filter producing updated estimation of the state $\hat{\mathbf{s}}_j(k|k)$, the likelihoods of the filters $\mathbf{\Lambda}_j(k)$ and the model probabilities $\mu_j(k)$. However, not all measurements originate from speakers. False measurements originate mainly due to reverberation and low energy segments. A validation region for each mode *j* at time k is constructed around the measurements based on the following predictions:

$$\hat{\mathbf{s}}_j(k|k\text{-}1)=\mathbf{F}\hat{\mathbf{s}}_j(k\text{-}1|k\text{-}1) \quad (5)$$

$$\hat{\mathbf{y}}_j(k|k\text{-}1)=\mathbf{H}\hat{\mathbf{s}}_j(k|k\text{-}1) \quad (6)$$

$$\mathbf{P}_j(k|k\text{-}1)=\mathbf{F}\mathbf{P}_{0j}(k\text{-}1|k\text{-}1)\mathbf{F}^T+\mathbf{Q}_j \quad (7)$$

New measurements are received from the array and validated if they lie within an acceptance region with probability P_G fulfilling:

$$e_j=(\mathbf{y}(k)-\hat{\mathbf{y}}_j(k|k\text{-}1))(\mathbf{S}_j(k))^{-1}(\mathbf{y}(k)-\hat{\mathbf{y}}_j(k|k\text{-}1))\le g_j^2 \quad (8)$$

where $\mathbf{S}_j(k)$ is the covariance of the innovation and g_j^2 (known as the number of standard deviations of the gate) is determined by P_G as well as the dimension of the state based on a chi-square test with a 99% (or 99.9%) confidence region [7]. The purpose of (8) is to construct a validation region (known as 'gate') to eliminate unlikely measurement-to-track pairings. The measurements inside the gate are considered to be possibly originated from the true speaker so as to be associated to the previous estimates forming a track. Otherwise the measurements are rejected in order not to affect the estimation procedure. PDA computes the probability of each validated measurement being generated from the speaker, β, (as well as the probability that no measurement is obtained from the speaker, β_0).

$$\beta_j=\frac{e_j}{b+\sum_{l=1}^{m} e_l},\ \beta_0=\frac{b}{b+\sum_{l=1}^{m} e_l},\ b=\lambda\sqrt{\det(2\pi\mathbf{S})}\,\frac{1-P_D P_G}{P_D} \quad (9)$$

λ: density of the clutter , P_D : the detection probability [7], assuming the clutter is uniformly distributed within the gate. The Kalman gain is defined for each mode *j* by:

$$\mathbf{K}_j(k)=\mathbf{P}_j(k|k\text{-}1)\mathbf{H}^T(\mathbf{H}\mathbf{P}_j(k|k\text{-}1)\mathbf{H}^T+\mathbf{R}_j)^{-1} \quad (10)$$

The a-posteriori state estimate and covariance for mode *j* are:

$$\hat{\mathbf{s}}_j(k|k)=\hat{\mathbf{s}}_j(k|k\text{-}1)\mathbf{K}_j(k)\sum_j\beta_j(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k\text{-}1)) \quad (11)$$

$$\mathbf{P}_j(k|k)=\beta_0\mathbf{P}_j(k|k)+(1-\beta_0)(\mathbf{I}-\mathbf{K}_j(k)\mathbf{H})\mathbf{P}(k|k\text{-}1)+\mathbf{P}^E \quad (12)$$

$$\mathbf{P}^E=\mathbf{K}_j(k)\sum_j\beta_j(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k\text{-}1))(\mathbf{y}(k)-\mathbf{H}\hat{\mathbf{s}}_j(k|k\text{-}1))^T\mathbf{K}_j(k)^T \quad (13)$$

Step 3 (Probability Update): The model probabilities are updated according to

$$\mu_j(k)=\frac{\Lambda_j(k)C_j(k-1)}{C} \text{ where } C=\sum_j \Lambda_j(k)C_j(k-1)\cdot$$

Step 4 (Output mixing): The final predicted state and covariances are computed by combining and weighting the estimates of all possible modes:

$$\hat{\mathbf{s}}(k|k)=\sum_j\mu_j(k)\hat{\mathbf{s}}_j(k|k) \quad (14)$$

$$\mathbf{P}(k|k)=\sum_j\mu_j(k)[\mathbf{P}_j(k|k)+[\hat{\mathbf{s}}_j(k|k)-\hat{\mathbf{s}}(k|k)][\hat{\mathbf{s}}_j(k|k)-\hat{\mathbf{s}}(k|k)]^T] \quad (15)$$

The algorithm iterates through step 1.

The corresponding state vector at time k for this case is $\mathbf{s}(k)=[\theta(k)\ \theta'(k)]^T$, and the measured position is $\mathbf{y}(k)=[\theta(k)]$ and the corresponding matrices in (1), (2) are:

$$F=\begin{bmatrix}1 & T\\ 0 & 1\end{bmatrix},\ Q_j=\begin{bmatrix}\frac{T^4}{4} & \frac{T^3}{2}\\ \frac{T^3}{2} & T^2\end{bmatrix}q_j,\ p_{ij}=\begin{bmatrix}p_{11} & p_{12}\\ p_{21} & p_{22}\end{bmatrix}=\begin{bmatrix}0.6 & 0.4\\ 0.4 & 0.6\end{bmatrix}$$

and T= block_size*overlap/sampling freq. We used 256 samples FFT at 8 kHz sampling rate with 50% overlap.

The IMM we used had two modes of movement. We have fixed the design parameters so that the non-moving and slowly moving mode possess low-level process noise (q_1=0.001) while the turning mode (manoeuvering model) possesses a much higher noise level (q_2=1).

4. THE GATE AS A VOICE ACTIVITY DETECTOR

A speaker can talk and then be silent for a long period making measurement-to-tracks association difficult. To deal with this problem we rely on the fact that in the context of DOA estimation a Kalman-based tracking algorithm incorporates source motion into angle estimates and can be characterized as an angle predictor followed by an observation-dependent angle corrector. The DOAs originating from a certain speaker cannot change randomly since the speaker's movement follows the Newtonian dynamics of motion. Using the previously estimated angle and angle rate the tracker forms an area where the next DOA is possible to lay. The permissible area is a circle centred at the predicted DOA having a radius proportional to the square root of the covariance matrix. Equation (8) that describes the formation of a validation gate around the predicted DOAs serves also as a means for the initiation and termination of the tracks of each individual speaker. If a number of consecutive measurements fall outside the validation gate the track is terminated. Track termination entails blocking of the microphone array from the direction of the speaker whose track is terminated. Subsequently, the initiation of both track and IMM filter is based on nulling the initial velocity and validating the new angle against the gate. Consistent rejection of DOA measurements for a speaker indicates that the particular speaker is silent. Therefore, the role of the gate is twofold:

a) Initiation and termination of DOA tracks (see Fig. 2).

b) Act as VAD for each speaker (see Fig. 3).

If an utterance contains long pauses or silence parts, the gating process will dissect the stream into disjoint speech segments. In the case of static speakers the disjoint segments can be associated with the same speaker. A similar association with moving speakers is not generally possible if we rely solely on the acoustic modality. The latter was anticipated as, for example, in the case of two speakers that would resume talking after switching (silently) their location. The latter switching process would be transparent to the acoustic modality regardless of the tracking or beamforming algorithm. However, in practice what is crucial for communication applications (i.e., speech or speaker recognition) is to have independent voice streams with as little competing voice interference as possible and to leave the speech or speaker recognition engine to deal with the interpretation of the voice streams. The proposed method relies on the robust initiation of tracks. The initiation of tracks is based on clustering the DOAs derived from the initial frames (250 ms).

We evaluated the proposed speech separation technique on a multispeaker speaker scenario taking place in 5x3x3 enclosure, where the utterances of the speaker contain large silence parts in between. We performed 10 simulations of the same experimental settings using concatenated digit signals randomly chosen from the NOISEX database and averaged the SIR results using the definition of SIR as in [10]. For each voice the rest were considered as interference. Since there are many amplification combinations by mean of which the signals reach the range of the input SIR of Table 1, we used the same amplification coefficient for each interference signal to reach the desired SIR and re-executed each simulation. The room layout is depicted in Fig. 1. The initial DOAs of the speakers are calculated from clustering the initial DOAs.

The SIR results are depicted in Table 1. We observe large improvement in the separation results, which is due to two factors: a) the directivity of the reception lobe and, b) the inactivation of the microphones from the direction of the silent speaker. The latter is achieved by nulling the voice stream of the inactive speaker as long as its corresponding DOAs do not fall in the predicted gate, therefore, preventing the interfering voices of the other speakers to leak in the stream of any inactive speaker. The functioning of the gating process as a VAD is illustrated in Fig. 3 and requires a buffer of three frames each of 256 samples at 8 kHz sampling frequency. Recordings of the separation results are included in http://slt.wcl.ee.upatras.gr/potamitis/IMMPDA_SMC.zip

SIR_In1 (dB)	SIR_ Imp_1	SIR_ Imp_2	SIR_ Imp_3
-10	14.57	28.34	20.65
-5	12.78	27.69	18.45
0	11.31	27.55	17.21
5	8.73	25.82	16.99
10	6.57	23.62	15.66

Table 1. Three static speakers, Signal to Interference Ratio (SIR) improvement. SIR_In1 is the SIR considering speaker 1 as the target speaker and the rest of the speakers as interference. SIR_Imp_(*i*) is the SIR improvement on speaker *i* considering the rest as interference.

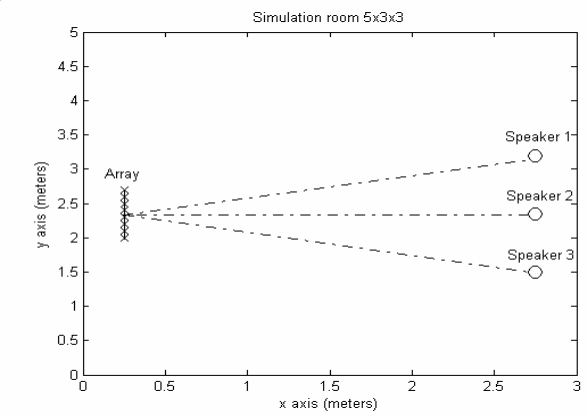

Fig. 1. Three static speakers in the enclosure.

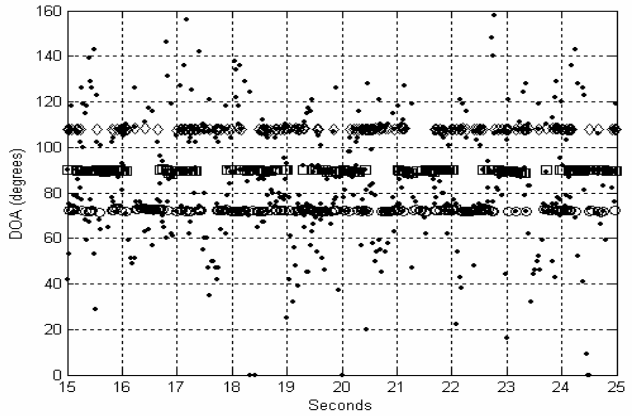

Fig. 2. Three partially overlapping speakers uttering random numbers. Distinct DOA clusters corresponding to words. (.) : DOA observations, (◊): Estimated DOAs associated to speaker 1, (□): Estimated DOAs associated to speaker 2, (○): DOAs associated to speaker 3.

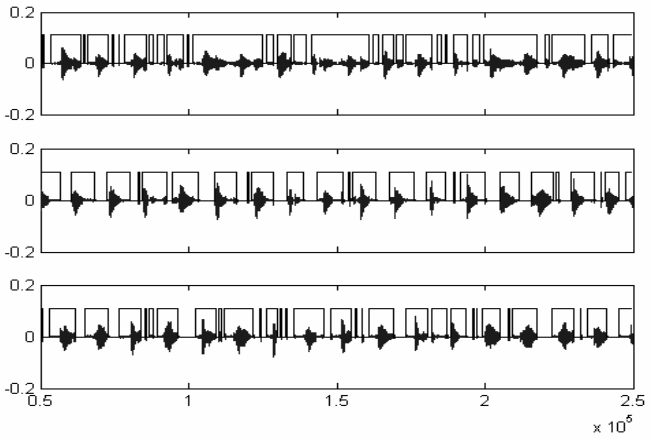

Fig. 3. The function of the gate functioning as a VAD. Three partially overlapping speakers uttering random numbers. The VAD is based on the DOA measurements falling inside the gate of each speaker.

5. MULTI-TARGET MULTI-SENSOR TECHNIQUES FOR VOICE SEPARATION AND EXTENSIONS

Much work has to be done in order to apply multi-target tracking techniques (mostly developed for radars that use simple narrowband signals) to take into account the idiosyncrasies of the speech signal which is a broadband, non-stationary signal. As there are distinct differences between speakers and moving targets, human motion and conversational attitude need to be taken into account; in polite conversations most speakers do not speak simultaneously, while target signals coexist for most of the tracking time. A speaker can talk and then be silent for a long period making hard the association of distinct track segments. However, the proposed method can derive independent streams corresponding to the different voices. Currently we are experimenting on the incorporation of a speaker recognition module that calculates the probability that a segment belongs to a specific speaker and to incorporate it in (9) to perform track initiation, termination and measurement-to-track association for complicated human interaction scenarios involving multiple speakers.

Although the versatility of human motion and the variability of the environment make quantitative analysis of speaker tracking problematic, extensive experimentation using the IMM filter strongly indicated that it is adequate to track any kind of speaker's motion. The use of more advanced data association techniques more suitable for the multi-speaker scenario as joint probabilistic data association or multiple hypothesis tracking [7] is expected to be beneficial as well as the use of heterogeneous sensors.

6. CONCLUSIONS

The hybrid state estimation with PDA to account for measurement origin uncertainty has the potential to unify the spatial sound selectivity and tracking the angles of arrival of multiple speakers using a single microphone array, thus providing a consistent and coherent way to reduce audio drop-out due to misaim. The proposed method provides independent voice streams in the case of partially overlapping, static-speakers case. We have also demonstrated how the prediction of IMM filters forms a gate of permissible DOAs that can be used to initiate and terminate a track, and therefore to segment independent streams to words.

REFERENCES

[1] Johnson D., Dudgeon D., "Array Signal Processing: Concepts and Techniques," *Prentice Hall*, 1993.

[2] Krim H., Viberg M., "Two Decades of Array Signal Proc. Research," *IEEE Signal Processing Magazine*, pp. 67-93, 1996.

[3] Kellerman W., "A self-steering digital microphone array," *Proc. IEEE ICASSP*, pp. 3581-3584, 1991.

[4] Huang Y., Benesty J., Elko G., Mersereau R., "Real-time passive source localization: an unbiased linear-correction least-squares approach", *IEEE Trans.on Speech &Audio Proc.*, vol. 9, no. 8, pp. 943-956, 2001.

[5] Yamada T., Nakamura S., Shikano K., "Distant-talking speech recognition based on a 3-D Viterbi search using a microphone array," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 2, pp. 48-56, 2002.

[6] Brandstein M., Silverman H., "A practical methodology for speech source localization with microphone Arrays," *Computer Speech and Language*, Vol. 2, pp. 91-126, 1997.

[7] Blackman S., Popoli R., "Design and analysis of modern tracking systems," *Artech House*, 1999.

[8] Potamitis I., Tremoulis G., Fakotakis N., "Multi-Speaker DOA Tracking Using Interactive Multiple Models and Data Association Techniques," *Proc. of EUROSPEECH*, Vol. I, pp. 517-520, 2003.

[9] Bar-Shalom Y., Li X., Kirubarajan T., "Estimation with application to tracking and navigation," *Wiley*, 2001.

[10] Mukai R., Sawada H., Araki S., Makino S., "Robust real-time blind source separation for moving speakers in a room," *IEEE Proc. of ICASSP*, Vol. 5, pp. 469-473, 2003.