Source Localization for Multiple Speech Sources Using Low Complexity Non-Parametric Source Separation and Clustering

M. Swartling[a,*], B. Sällberg[a], N. Grbić[a]

[a] *Department of Electrical Engineering, Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden*

Abstract

This article presents a new method for localization of multiple concurrent speech sources that relies on simultaneous blind signal separation and direction of arrival (DOA) estimation, as well as a method to solve the intersection point selection problem that arises when locating multiple speech sources using multiple sensor arrays. The proposed method is based on a low complexity non-parametric blind signal separation method, making is suitable for real-time applications on embedded platforms. On top of reduced complexity in comparison to a previously presented method, the DOA estimation accuracy is also improved. Evaluation of the performance is done with both real recording and simulations, and a real-time prototype of the proposed method has been implemented on a DSP platform to evaluate the computational and the memory complexities in a real application.

Keywords: source localization, direction of arrival estimation, acoustic arrays, speech processing

1. Introduction

Traditional blind signal separation methods have been designed and used for signal separation aiming at reconstruction or estimation of original source

*Corresponding author, tel. +46 455 385581, fax +46 455 385057

Email addresses: msw@bth.se (M. Swartling), bsa@bth.se (B. Sällberg), ngr@bth.se (N. Grbić)

signals, and to maintain the audio quality of the reconstructed source signals. One approach to blind speech signal separation is to exploit the sparse and disjoint structure of the signals in the time-frequency domain. It has been shown that masking of time-frequency points can separate additive mixtures of speech signals [1]. Various methods to design the masks have been presented, including methods based on signal model mixing parameters, observation vectors or direction of arrival estimates, and independent components [1, 2, 3, 4, 5]. Masking in the time-frequency domain generally introduces music distortion in the reconstructed signals, and there is often a tradeoff between the amount of separation and the level of introduced music distortion. Research aiming to reduce the music distortion through masking while maintaining a good signal separation has been conducted [6, 7].

For the purpose of estimating the DOA of the separated sources, the signals do not necessarily have to be reconstructed. Any audible music distortion that would have appeared in the reconstructed signals due to masking does not have to be a concern either. The previously presented method [8] used a blind signal separation method based on DUET [1] with the aim of estimating the DOA of the separated sources instead of reconstructing the source signals. It was possible to tune the parameters of the blind signal separation stage for more aggressive and faster converging separation, and the DOA of each source was then estimated by using the robust steered response power (SRP) with phase transform weighting (PHAT) method. Even though the separated speech sources empirically had a significantly lower perceptual quality due to the aggressive separation when the signals were reconstructed to the time domain, the DOA estimation performed well.

Accurate information about the location of sources is of great benefit to many applications in signal processing. Blind methods, or automatic tracking methods, for example, can sometimes take advantage of this additional knowledge to improve their performance. This includes applications such as speech enhancement and separation in hand held and mobile devices, laptops and conference telephony systems, and also automatic camera tracking in video conference,

surveillance and security systems.

In this article, a new method for DOA estimation of multiple concurrent speech sources, as well as a method to solve the problem with multiple combinations of intersection points that arises when locating multiple speech sources using multiple sensor arrays, is presented. The combined stages provides a new method for positional location estimation of multiple concurrent speech sources using multiple small sensor arrays. The method previously presented in [8] performed the blind signal separation and DOA estimation in two separate stages: the first stage separated the sources with a conventional signal separation method based on DUET, and the second stage performed the DOA estimation. In this article, the new proposed method combines the two stages into a new simultaneous blind signal separation and DOA estimation method. The new proposed method eliminates the DUET for the signal separation stage from the previous method, and instead replaces it with a feedback from the DOA estimation stage. In addition to evaluating the performance with both real recordings and simulations, a real-time prototype of the proposed method has been implemented on a DSP platform to evaluate the computational and the memory complexities in a real application.

When employing multiple sensors arrays for source localization in the far field, intersecting DOA estimates from different sensor arrays yields the physical location of the source [9]. In the presence of multiple sources, there are multiple DOA estimates at each sensor array and also multiple combinations of intersection points, many of which do not correspond to real physical sources. Also proposed in this article is an addition to the method presented in [10] to handle the problem with multiple intersection points, which is solved by correlating parameters from the blind signal separation stage. This allows the localization stage to intersect the correct DOA estimates from multiple sensor arrays so the resulting position estimate corresponds to a true source location. Other methods to solve the multiple intersection point problem include methods based on clustering of intersection points and correlation of signals [11, 12].

Figure 1 shows the problem when intersecting DOA estimates for localiza-

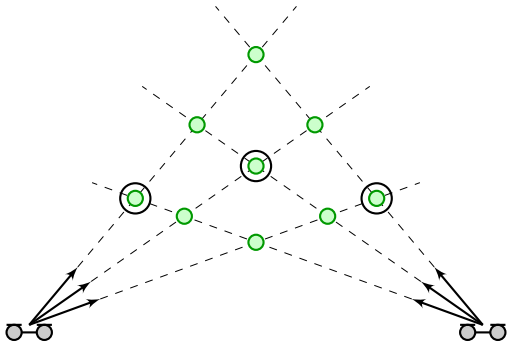

Figure 1: The problem when intersecting three DOA estimates at two sensor arrays. Arrows represent the estimated DOAs for each source at each sensor array, and the intersection points are shown as circles. Multiple intersection points exists, many of which do not correspond to real physical sources. The circles in the figure show one of six possible configurations of true intersection points.

tion of three sources with two sensor arrays. The solid arrows represent the estimated DOA vectors of the sources, and each sensor array estimates the DOA for each source. The intersection points are indicated by the circles. In the figure, there are nine intersection points and six possible ways to combine them to estimate the location of the sources. In the general case, for N sources, there are N^2 intersections and $N!$ combinations which might originate from real physical sources. The terms true and false intersection points are used to describe whether an intersection point corresponds to a real source location or not. A true intersection point is the result of pairing DOA estimates for the same source from different sensor arrays, while false intersections results from pairing vectors from different sources.

2. Multi-source DOA Estimation Stage

The first problem to address is the DOA estimation for multiple concurrent speech sources. A cluster-based blind signal separation method is proposed

using a time-frequency domain masking approach. The observed time delay estimates for each time-frequency point are clustered around a set of cluster centroids. After masking the time-frequency points into the separate clusters, the proposed method estimates the DOA to each source and adjusts the cluster centroids using the SRP-PHAT method.

2.1. W-disjoint Orthogonality

Common to the time-frequency masking-based methods is the assumption that the time-frequency representations of the signal components are sparse and that their supports are disjoint. Denoted as W-disjoint orthogonal [1], the methods exploit the sparse and disjoint structure of speech signals in the time-frequency domain for blind signal separation.

For two time signals $s_1(t)$ and $s_2(t)$, and the corresponding time-frequency representation $S_1(\omega, \tau)$ and $S_2(\omega, \tau)$, the disjointness of signals can be expressed as

$$S_1(\omega, \tau)S_2(\omega, \tau) = 0. \tag{1}$$

The orthogonality states that there is no energy overlap in the time-frequency domain. From the W-disjoint orthogonality it follows that a set of time-frequency masks exists that can separate an additive mixture into its components. Under the assumption that the sources are indeed W-disjoint orthogonal, at most one source contribute with energy at a specific time-frequency point (ω, τ), and a properly chosen mask will therefore mask the time-frequency point from a specific source. The disjointness of speech signals have been extensively researched, and it has been shown that multiple independent speech signals are nearly W-disjoint orthogonal [1].

2.2. Blind Signal Separation

The propagation path is assumed to be anechoic, where the sensor signals are subject to a time-shift and an attenuation due to the propagation distance. For a small sensor array, the attenuation is assumed to be equal all over the

sensor array, and is modeled by a unit gain. The sensor signal for the sensor m is modeled as

$$x_m(t) = \sum_{n=1}^{N} s_n(t - \delta_{n,m}) + \nu_m(t) \tag{2}$$

where N is the number of sources and is assumed to be known. The source signals $s_n(t)$ are subject only to a time-shift $\delta_{n,m}$ from the source n to the sensor m. The sensor noise $\nu_m(t)$ is modeled as zero-mean independent white Gaussian noise. It is assumed that the time-shift can, in the absence of noise, be modeled as a phase-shift in the time-frequency domain [13]:

$$x_2(t) = x_1(t - \delta) \leftrightarrow X_2(\omega, \tau) = X_1(\omega, \tau)e^{-j\omega\delta}. \tag{3}$$

In the discrete time-frequency domain, henceforth denoted by $[\omega, \tau]$ with discrete time τ and frequency ω, and with the propagation delay $\delta_{n,m}$ denoting samples, the signal model is

$$X_m[\omega, \tau] = \sum_{n=1}^{N} S_n[\omega, \tau]e^{-j\omega\delta_{n,m}} + N_m[\omega, \tau]. \tag{4}$$

Under the assumption that the sources are W-disjoint orthogonal, no more than one source is active at any time-frequency point. The signal model can thus be reduced to

$$X_m[\omega, \tau] = S_n[\omega, \tau]e^{-j\omega\delta_{n,m}} + N_m[\omega, \tau] \tag{5}$$

where n is the index of the single active source at the corresponding time-frequency point $[\omega, \tau]$. The cross-power spectrum for two sensor signals then becomes

$$G_{p,q}[\omega, \tau] = \mathrm{E}\left[X_p[\omega, \tau]X_q^*[\omega, \tau]\right] \tag{6}$$

$$= |S_n[\omega, \tau]|^2 \, e^{-j\omega(\delta_{n,q} - \delta_{n,p})} \tag{7}$$

where $\mathrm{E}[\cdot]$ is the expectation operator. For each time-frequency point, a time delay $T_{p,q}[\omega, \tau]$ between sensors p and q can be derived from the cross-power spectrum as

$$T_{p,q}[\omega, \tau] = \delta_{n,q} - \delta_{n,p} \tag{8}$$

$$= \frac{1}{\omega}\angle G_{p,q}[\omega, \tau]. \tag{9}$$

For a small uniform linear sensor array with more than two sensors, an average time delay for a time-frequency point can be calculated as the average of the time delays for all sensor pairs as

$$T[\omega, \tau] = \frac{1}{|\mathbf{P}|} \sum_{\{p,q\}} \frac{T_{p,q}[\omega, \tau]}{q - p}. \tag{10}$$

The set $\mathbf{P}$ contains the sensor pairs $\{p, q\}$ in the sensor array and $|\mathbf{P}|$ denotes its cardinality. Due to the assumed W-disjoint orthogonality, and the anechoic propagation path, the time delays depend only on what source is active in the given time-frequency point. Time delay estimates for all time-frequency points that are dominated by the same source will have the same value.

The estimated time delay for each time-frequency point is dependent on the phase of the cross-power spectrum. Thus, if the sensor array is too large and introduces spatial aliasing, the phase at some point in the cross-power spectrum becomes ambiguous. It is therefore necessary to ensure that there is no spatial aliasing by requiring that the inter-sensor distance d for any sensor pair $\{p, q\} \in \mathbf{P}$ is $d < c/(2 \cdot f_{\max})$, where c is the propagation speed of sound and $f_{\max}$ is the highest frequency in hertz. The sensor arrays used in the evaluation section have an inter-sensor spacing of $d = 0.04$ m, so $f_{\max} < 4250$ Hz. Thus, sampling at 8 kHz satisfies the limit for spatial aliasing.

A number of cluster centroids C_n, one for each source present in the received sensor signals, are used to track the clusters of time delays for each time-frequency point. From the cluster centroids, a binary time-frequency mask for the source n is calculated as

$$\mathbb{M}_n[\omega, \tau] = \begin{cases} 1 & T[\omega, \tau] \text{ is closest to cluster } C_n \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The original source n, as recorded by sensor m, is demixed in the time-frequency domain as

$$\hat{S}_{n,m}[\omega, \tau] = \mathbb{M}_n[\omega, \tau] X_m[\omega, \tau]. \tag{12}$$

For source reconstruction, any one of the received sensor signals can be demixed and the time-frequency representation is transformed back to time domain. For

source localization, which is the focus of this article, it is necessary to demix all sources from all sensor signals.

2.3. Time Delay Estimation

Since the original sources have been separated, conventional single-source DOA estimation methods can be used to locate the individual sources. A popular method for localization is the SRP-PHAT method, which is based on maximizing the output power from a delay-and-sum beamformer by steering it across a defined search space. The output power from the delay-and-sum beamformer for source n is

$$P_n(\tau) = \sum_{\{p,q\}} \sum_{\omega} \psi[\omega] G_{n,p,q}[\omega] e^{j\omega\tau(q-p)} \tag{13}$$

where $\psi[\omega]$ is a weighting function. A number of weighting functions are presented in [14]. One of the more popular weighting functions is the phase transform (PHAT):

$$\Psi_{\mathrm{PHAT}}[\omega] = \frac{1}{|G_{n,p,q}[\omega]|}. \tag{14}$$

The phase transform has been shown to perform well for speech signals in reverberant environments [15, 16]. The acoustic source is then located by maximizing the output power of the beamformer:

$$\hat{\tau}_n = \arg\max_{\tau} P_n(\tau). \tag{15}$$

The signal model and the W-disjoint orthogonality assumes that at most a single source is active in any time-frequency point. The cross-power spectrum can therefore be estimated not in terms of the sensor signals, but in terms of the separated source signals:

$$\hat{G}_{n,p,q}[\omega] = \hat{\mathrm{E}}\left[\hat{S}_{n,p}[\omega,\tau]\hat{S}^*_{n,q}[\omega,\tau]\right] \tag{16}$$

where $\hat{\mathrm{E}}[\cdot]$ is the sample-based estimator of the expected value. The signals $\hat{S}_{n,p}[\omega,\tau]$ and $\hat{S}_{n,q}[\omega,\tau]$ are the estimates of the signal n in time-frequency point $[\omega,\tau]$ as received by the sensors p and q, respectively.

3. Proposed Cluster Centroid Tracking Stage

The proposed method is a novel method for updating cluster centroids in an online centroid tracking environment aiming at estimating the DOA of the separated sources. The previously presented method in [8] consists of two separate stages for estimating the DOA for multiple speech sources. The first stage was based on real-time DUET [17], in which the signal model mixing parameters are tracked online using a gradient-based search method of a maximum likelihood cost function. The tracked mixing parameters serve as cluster centroids for the observed mixing parameters, and are used to create the masks to separate the individual sources. The second stage estimates the DOA of the separated signals using the robust SRP-PHAT method. In the new proposed method, the two stages are combined and the DOA information from the robust SRP-PHAT is fed back to the cluster centroid tracking stage. The DUET is thus eliminated from the signal separation stage and replaced with the proposed cluster centroid tracking stage.

The cluster centroids track the time delay signal mixing parameter. The observed feature vectors for clustering are the observed time delays $T[\omega, \tau]$ for each time-frequency point $[\omega, \tau]$. After separation, the SRP-PHAT method updates the cluster centroids. The estimated TDOA $\hat{\tau}_n$ for the source n from the SRP-PHAT equals the new cluster centroid, so the cluster update is $C_n = \hat{\tau}_n$.

Figure 2 illustrates the recursive process for signal separation and TDOA estimation. The clustering and masking stages makes the signal separation stage, and the TDOA estimation stage makes the cluster update stage.

4. Proposed Intersection Point Selection Stage

By separating the sources and estimating the DOA at multiple sensor arrays, the DOA estimates can be intersected for localization purposes. The intersection point selection problem consists of intersecting the correct DOA estimates from the sensor arrays. A pair of masks produced by the blind signal separation stage at different sensor arrays for the same source are assumed to be correlated.

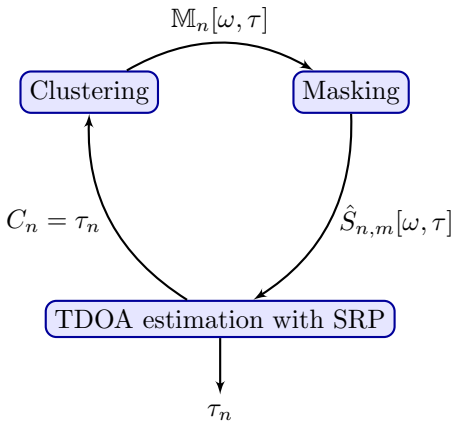

Figure 2: The recursive process for signal separation and TDOA estimation for multiple sources.

The proposed methods to match signals and DOA estimates between arrays are based on correlating the masks used by the separation and the separated signals. Masks and signal estimates that correlate are assumed to belong to the same source, and so the corresponding DOA estimates intersect at a real physical source location.

By separating the sources at each sensor array, a set of masks $\mathbb{M}_n^k$ is estimated for each of the sources n at each sensor array k. A straightforward way to select a combination of DOA pairs from two sensor arrays is to calculate an $N \times N$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & a_{N,N} \end{bmatrix} \tag{17}$$

where

$$a_{n_1,n_2} = \left\| \mathbb{M}_{n_1}^1 \odot \mathbb{M}_{n_2}^2 \right\|. \tag{18}$$

and where $\odot$ denotes element-wise multiplication and $\|\cdot\|$ denotes the Euclidean norm. A larger value of a matrix element a_{n_1,n_2} means that there are many common time-frequency points, which implies a higher probability that the corresponding DOA pair intersects at a true source location.

To estimate a set of DOA pairs that intersect at true source locations, successively select the largest element from the matrix $\mathbf{A}$, while the corresponding row and column should be excluded from the following iterations. The row and column indices correspond to the DOA indices for the corresponding sensor array. By excluding the selected row and column indices from the following iterations, it is ensured that each DOA estimate is used exactly once in the estimation of a source location. This approach assumes that there is only one source for each DOA vector. However, this is a valid assumption, since the blind signal separation stage separates the sensor signals into the individual source signals.

When correlating the binary masks, the energy contained in each time-frequency point of the separated signals is not weighted into the mask with the result that noise will contribute to the correlation estimates. To weight the correlation contribution for each time-frequency point, the masks can be weighted by the energy content in the corresponding time-frequency points. The elements of the matrix then becomes

$$a_{n_1,n_2} = \|\hat{S}^1_{n_1,m} \odot \hat{S}^2_{n_2,m}\|. \tag{19}$$

Since the separated signals $\hat{S}_{n,m}$ are already masked during the separation of the sources from the mixtures, the energy-weighted masks equals the energy of the separated signals. Similar to the demixing in (12), any one of the demixed sensor signals can be used.

Finding many common time-frequency points does not only imply a good DOA match, but it also suggests that a higher number of common dominant time-frequency points is used for the DOA estimation. To exploit this assumption, a scoring system is also implemented. Different scoring systems based on the matrix $\mathbf{A}$ and the corresponding sensor time-frequency signals were investigated. However, the scoring system that turned out to work best was to directly use the value of a_{n_1,n_2} as the score for the pair (n_1, n_2). An additional observation regarding the proposed method is the order in which the pairs are selected. The pairs are selected from $\mathbf{A}$ in decreasing order of the corresponding score

a_{n_1,n_2}. The number of selected permutation pairs may therefore be limited to only the number of pairs with the highest score. That means that only the intersection points with the highest correlation are accepted.

The difference in the two proposed scoring systems is what mutual information between the signals that is emphasized. The score for correlating the binary masks is a measure of the number of common dominant time-frequency points. However, for time-frequency points where none of the signals contain a significant amount of energy, the mask is noisy, which contributes equally to the final score. When correlating the weighted masks, the score is almost exclusively determined by the correlation of the high energy formants of the speech signals. Even a single erroneously masked formant can then significantly alter the score in favor of false intersections.

5. Evaluation

The evaluation is performed in both a simulated room environment and using real room recordings. The blind signal separation and the DOA estimation stages are evaluated using simulated sensor array data to cover a range of controlled environments. Real room recordings are used to evaluate the intersection point selection stage to show its practical use. Room impulse responses for simulation are generated by the image method [18], where reflection coefficients for the room boundaries are adjusted for different reverberation times. Sensor array data are finally generated by filtering a source signal by the generated synthetic room impulse responses.

The real sensor data are recorded in a typical small conference room environment measuring $4 \times 4 \times 2.6$ meter. The same room dimensions and sensor and source placements are modeled to generate the synthetic impulse responses. The two sensor arrays consist of four sensors each, with 4 cm sensor spacing. The sample rate is 8 kHz, and the time-frequency transformation of the sensor signals is performed using oversampled uniform DFT filterbanks [19, 20]. The number of subbands vary, but throughout the evaluations, a two times oversam-

pling and two-tap polyphase subband filter, have been used. Speech sequences, both for simulation and the real recordings, are taken randomly from the TIMIT database.

5.1. Multi-Source DOA Estimation

The blind signal separation is evaluated using simulated room impulse responses. Two sources are placed at a distance of 1.5 meter from a single sensor array: one source is placed at $-30°$ and a second source at $0°$, $20°$, $40°$ and $60°$, from the sensor array broadside. The proposed SRP-based method is compared to the previous DUET-based multi-source DOA estimation method for a varying number of subbands, reverberation times and angular source spreading.

The evaluation results are presented in table 1. Estimates of the DOA for each of the two sources are produced at every sample period of the analysis filterbank output. For 64, 256 and 1024 subbands and an oversampling factor of 2, the DOA estimates are produced every 4, 16 and 64 millisecond, respectively. The average standard deviation and mean estimation errors over the different positional configurations are presented in the table. The mean estimation error is calculated as $\text{MEE} = \hat{\text{E}}\left[|\alpha| - |\hat{\alpha}|\right]$, where $\hat{\text{E}}\left[\cdot\right]$ is the sample based expectation operator, and represents the mean error in the estimated DOA compared to the true DOA. A positive estimation error is an offset towards the array's endfires. As shown in the table, the proposed method performs with a lower standard deviation and mean estimation error than the previous DUET-based method in all setups.

The results show a general trend towards larger standard deviation and mean estimation error when the number of subbands is reduced. However, the number of subbands also controls the delay of the filterbanks and is a tradeoff between signal separation and DOA estimation performance, and the delay between the input and output signals. The delay can be reduced by decreasing the polyphase subband filter length at the expense of increased subband aliasing, or by increasing the oversampling at the expense of a higher subband sample rate and higher processing requirements. The filterbank structure can furthermore be restricted

Table 1: Evaluation of the proposed SRP-based method in comparison to the previous DUET-based method for multi-source DOA estimation. The estimated DOA is evaluated for standard deviation (STD) and mean estimation error (MEE). The results are averaged over five speech recordings and four positional configurations.

(a) $RT_{60} = 100$ ms

	SRP		DUET	
Subbands	STD	MEE	STD	MEE
64	2.37	-4.08	5.01	-6.51
256	1.04	-1.97	1.82	-3.60
1024	0.65	-1.36	0.96	-2.79

(b) $RT_{60} = 250$ ms

	SRP		DUET	
Subbands	STD	MEE	STD	MEE
64	3.48	-7.34	5.05	-13.30
256	1.53	-6.60	2.67	-10.93
1024	0.77	-6.34	1.28	-10.09

by other processing, such as speech enhancement, if the same filterbanks are utilized.

5.2. The Intersection Point Selection Problem

The intersection point selection method is evaluated with simulations as well as real room recordings. The real recording setup consists of two sensor arrays and three sources, where the sensor array geometries and filterbank structures are the same as in the simulation setup for multi-source DOA estimation. Filterbanks with 256 subbands are used.

Figure 3 shows a part of a single recording and corresponding localization

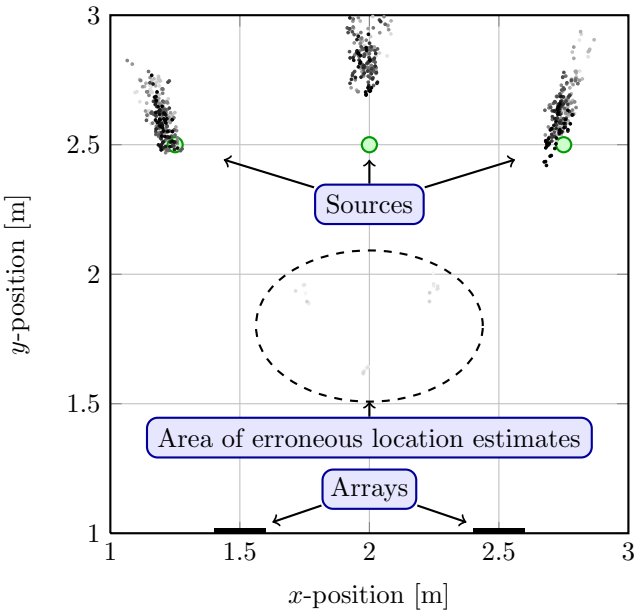

Figure 3: An overhead view of the room with three sources, where each dot is a location estimate at each time instant. Location estimates are scored according to the correlation of the binary masks in (18). Source DOA estimation and DOA pair selection are performed with the proposed methods using filterbanks with 256 subbands. The dashed circle marks the area where most of the erroneous location estimates are clustered.

results. The graphs represent an overhead view of the room, and each dot is a location estimate at each time instant. A location estimate for each of the three sources is produced at every sample period of the analysis filterbank output. Each location estimate is colored based on the score, where a darker color is an estimate with a higher score. The true source locations are shown as circles, and the two sensor arrays are shown as black bars in the bottom of the figure.

Some locations due to erroneously selected DOA pairs are found, mostly visible between the sensor arrays and true source locations, as indicated in figure 3 by the dashed circle. These erroneous location estimates, however, have a lighter color and thus have a lower score. The location estimates with

Table 2: The percentage of correctly matched DOA pairs for three to five sources. The "no weight" column is the percentage of correctly matched pairs, and the "score weight" column is the percentage when weighted by the corresponding score in (18).

		Pairs correctly matched [%]	
	Score by	No weight	Score weight
3 sources	Mask	97	98
Recorded	Signal	89	95
3 sources	Mask	95	98
Simulated	Signal	96	99
4 sources	Mask	94	98
Simulated	Signal	96	99
5 sources	Mask	90	92
Simulated	Signal	81	84

a higher score are mainly clustered near the true source locations. Note also how the clusters, most notably the center cluster, are not centered around the true source positions as indicated by the circles. The DOA estimates are shifted towards the broadside of each sensor array, consistent with the mean estimation errors in the simulated results in table 1.

Table 2 shows the percentage of correctly matched DOA pairs over many recordings with three sources from a real recording of varying speech phrases, and up to five sources for the simulated environment. A DOA pair is considered to be correctly chosen when the resulting intersection point is closer to any of the true intersection points than to any of the false intersection points. The two proposed score methods have been evaluated: the correlation of the masks, and the correlation of the separated source signals. Furthermore, two weighting methods have been evaluated: no weighting which corresponds to the percentage

of correctly matched DOA pairs, and score weighting which is the percentage weighted by the score of each location estimate. It can be seen from the table that a success rate of over 95%, both with and without score weight, is achieved in the majority of the test cases. Only the case with five sources shows a general decrease in success rate. The masks and the separated signals are sufficiently similar between sensor arrays, and can thus be used to pair DOA vectors that belongs to the same source.

6. Implementation Details

6.1. Computational Complexity

Common steps in both the presented SRP-based method and the previous DUET-based method is the masking of the sensor signals to separate the original sources, estimation of cross-power spectrums (CPS) and the DOA estimation using SRP-PHAT. The difference is how the cluster centroids are updated and how the masks are calculated. Table 3 lists the approximate number of operations needed to perform the steps that make the difference between the two methods. The listed operations are: complex valued additions and subtractions, multiplications and divisions, exponential and trigonometric functions, and search for a minimum value among a set of values. The number of operations is listed for each subband and time-step.

Table 4 lists the approximate number of operations needed to estimate the cross-power spectrum (CPS) and to evaluate the SRP-PHAT in (13) for a particular time delay τ. The cross-power spectrum is estimated using a first order AR-process, making the computational complexity low and independent of the effective integration time for the estimated value. The number of times needed to evaluate (13) depends on the numerical search method employed to find the point of maximum value in (15). The search method used during evaluation was a two-step approach. First, a coarse search across the search space at evenly distributed time delays τ was performed, followed by a golden section search with parabolic interpolation [21] around the initial highest steered response power

Table 3: Complexity comparison of the differences between the proposed SRP-based method and the previous DUET-based method. The complexity is an approximate calculation of the number of operations performed for each subband and time instant.

Operation	SRP	DUET
Addition	$\|\mathbf{P}\| + N$	$\|\mathbf{P}\| \cdot N \cdot 3 + N$
Multiplication	0	$\|\mathbf{P}\| \cdot N \cdot 9 + N$
Trigonometry	1	$\|\mathbf{P}\| \cdot N \cdot 4$
Minimum	1 (out of N)	1 (out of $\|\mathbf{P}\| \cdot N$)

to narrow in on the peak. The point of maximum steered response power was generally found in 18 to 25 evaluations.

The complexity of the proposed SRP-based method in comparison to the previous DUET-based method is reduced by an order of magnitude. However, the total complexity is dominated by the SRP-PHAT and the search for maximum steered response power. The total complexity improvement by the proposed method is thus reduced. Table 5 shows the complexity improvements for a varying number of sensors and sources. The complexity improvements are calculated from the approximated complexities in table 3 and 4, where M is the number of sensors in an array and N is the number of sources. Sensor pairs are selected as adjacent sensors within the array, so $|\mathbf{P}| = M - 1$, and the SRP-PHAT is assumed to be evaluated 20 times. The total complexity is thus reduced by 12–13%. If the search method is improved such that the number of SRP-PHAT evaluations are reduced, the total reduction by using the proposed method is increased. As an example, if the number of SRP-PHAT evaluations is reduced to 10, the improvement in table 5 is increased to approximately 25%.

6.2. Implementation

The proposed method has been designed with focus on realtime online processing. A prototype implementation has been implemented on a custom float-

Table 4: Complexity listing of the common steps; the cross-power spectrum (CPS) estimation and the SRP-PHAT evaluation. The complexity is an approximate calculation of the number of operations performed for each subband and time instant.

Operation	CPS estimation	+ SRP-PHAT
Addition	$\|\mathbf{P}\| \cdot N \cdot 2$	$+ \|\mathbf{P}\| \cdot N + \|\mathbf{P}\|$
Multiplication	$\|\mathbf{P}\| \cdot N \cdot 3 + \|\mathbf{P}\| + M$	$+ \|\mathbf{P}\| \cdot N \cdot 2$
Trigonometry	0	$+ \|\mathbf{P}\| \cdot N \cdot 2$

Table 5: Complexity improvement for the proposed SRP-based method over the previous DUET-based method. The number of operations are calculated from table 3 and 4 for different number of sensors M and sources N.

Configuration	SRP	DUET	Improvement
$M = 2$ and $N = 2$	239	271	12%
$M = 2$ and $N = 8$	881	1015	13%
$M = 8$ and $N = 2$	1637	1867	12%
$M = 8$ and $N = 8$	6059	7003	13%

ing point DSP platform. The two most prominent factors that could limit an implementation on the platform are the computational complexity and the memory complexity. The computational complexity was evaluated by comparing the frame time and the processing time for each of the stages. The frame time is the sample time of the output of the filter banks used to transform the sensor signals into the time-frequency domain. The memory complexity was evaluated as the persistent states for each of the stages and the maximum temporary working set.

An immediate observation and a critical factor during the implementation is the result of the signal separation stage. A cross power spectrum is estimated for each source and for each sensor so the DOA can be estimated for each separated source. An approach to reduce the memory impact of the signal separation stage is to assume that the set of sensor pairs $\{p, q\}$ for DOA estimation consists only of pairs of adjacent sensors. The memory impact can then be significantly reduced by only keeping one cross power spectrum for each source in memory. The memory impact is also reduced by using an AR-process to integrate the cross power spectrums. A variable integration time is achieved at a low constant memory impact. Another limiting factor is the number of filterbank subbands. Better localization performance can be achieved with a larger number of subbands, but at the cost of an increased memory impact. While the number of subbands increases the computational complexity per frame, the frame time is increased accordingly and the relative computational complexity remains constant.

Table 6 shows the computational and the memory complexity of the prototype implementation. The computational complexity is listed as the percentage of time spent in the stage relative to the frame time. The memory complexity is the storage requirements for persistent states. For as many as five concurrent sources being separated and located, the total computational complexity is 10.9% of the available processing cycles, and the memory complexity is 14.1% of the available memory. The sampling and filterbank setup is otherwise the same as in the evaluation section: 8 kHz sampling rate, 256 subbands, two times

Table 6: Computational and memory complexity for the stages.

Stage	Computations [$\times 10^6$ cycles]	Memory [$\times$32 bits]
Filterbanks	2.2	2048
Signal separation	0.8	5632
DOA estimation	7.2	10
Intersection selection	0.7	—
Maximum working set	—	1536
Total	10.9	9226

oversampling and two-tap polyphase subband filters.

7. Conclusions

The article presents a new method for simultaneous blind signal separation and DOA estimation of multiple speech sources, as well as a method to solve the intersection point selection problem when locating the sources using multiple sensor arrays. The methods integrate well, and form a complete solution to estimate the location of multiple speech sources. The improvements over the previously presented DUET-based method is the elimination of any signal or environment-dependent tuning parameters, and a significantly reduced computational complexity. The proposed method performs overall better for a wide range of simulated environments, and real room recordings show the method's practical use.

Two scoring methods are proposed and evaluated. One method is based on correlating the binary masks and one is based on correlating the energy of the separated source signals. The difference in the two methods is what mutual information is emphasized. The correlation of the binary masks is robust to estimation errors of the speech formants, while the correlation of the energy of the separated source signals is robust to mask noise. Both presented methods work

well in low-noise environments, where more than 95% or all location estimates are situated near real physical source locations.

The most computationally demanding stage is the SRP-PHAT step conducted to search for the peak in the steered response power from a sensor array. While most overall improvements can be made by reducing the number of evaluations of the steered response power, the proposed method offers a significant computational reduction over the DUET-based method. Over all, the computational complexity is improved by 12—13%. A prototype implementation on an embedded DSP platform demonstrates that, for as many as five concurrent sources, the computational and memory complexity is well within bounds of the platform.

References

[1] Ö. Yılmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Transactions on Signal Processing 52 (7) (2004) 1830–1847.

[2] S. Araki, H. Sawada, R. Mukai, S. Makino, A novel blind source separation method with observation vector clustering, in: Proc. International Workshop on Acoustic Echo and Noise Control, 2005.

[3] J. Cermak, S. Araki, H. Sawada, S. Makino, Blind speech separation by combining beamformers and a time frequency binary mask, in: Proc. International Workshop on Acoustic Echo and Noise Control, 2006.

[4] S. Araki, S. Makino, A. Blin, R. Mukai, H. Sawada, Underdetermined blind separation for speech in real environments with sparseness and ICA, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 3, 2004, pp. iii/881–iii/884.

[5] H. Sawada, S. Araki, R. Mukai, S. Makino, Blind extraction of dominant target sources using ICA and time-frequency masking, IEEE Transactions on Audio, Speech, and Language Processing 14 (6) (2006) 2165–2173.

[6] S. Araki, S. Makino, H. Sawada, R. Mukai, Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 3, 2005, pp. iii/81–iii/84.

[7] S. Araki, H. Sawada, R. Mukai, S. Makino, Blind sparse source separation with spatially smoothed time-frequency masking, in: Proc. International Workshop on Acoustic Echo and Noise Control, 2006.

[8] M. Swartling, N. Grbić, I. Claesson, Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 4, 2006, pp. 833–836.

[9] M. S. Brandstein, J. E. Adcock, H. F. Silverman, A closed-form location estimator for use with room environment microphone arrays, IEEE Transactions on Speech and Audio Processing 5 (1) (1997) 45–50.

[10] M. Swartling, M. Nilsson, N. Grbić, Distinguishing true and false source locations when localizing multiple concurrent speech sources, in: Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop, 2008, pp. 361–364.

[11] E. D. Di Claudio, R. Parisi, G. Orlandi, Multi-source localization in reverberant environments by ROOT-MUSIC and clustering, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 2, 2000, pp. 921–924.

[12] T. Nishiura, T. Yamada, S. Nakamura, K. Shikano, Localization of multiple sound sources based on a CSP analysis with a microphone array, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Vol. 2, 2000, pp. 1053–1056.

[13] R. Balan, J. Rosca, S. Rickard, J. O'Ruanaidh, The influence of windowing

of time delay estimates, in: Proc. Conference on Information Sciences and Systems, Vol. 1, 2000, pp. 15–17.

[14] C. H. Knapp, G. C. Carter, The generalized cross correlation method for estimation of time delay, IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-24 (4) (1976) 320–327.

[15] M. Brandstein, D. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

[16] C. Zhang, D. Florêncio, Z. Zhang, Why does PHAT work well in low noise, reverberative environments?, in: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, 2008, pp. 2565–2568.

[17] S. Rickard, R. Balan, J. Rosca, Real-time time-frequency based blind source separation, in: Proc. International Workshop on Independent Component Analysis and Blind Signal Separation, 2001, pp. 651–656.

[18] J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics, Journal of the Acoustical Society of America 65 (4) (1979) 943–950.

[19] P. P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice Hall, 1993.

[20] K. F. C. Yiu, N. Grbić, S. Nordholm, K. L. Teo, Multi-criteria design of oversampled uniform DFT filter banks, IEEE Signal Processing Letters 11 (6) (2004) 541–544.

[21] G. E. Forsythe, M. A. Malcolm, C. B. Moler, Computer Methods for Mathematical Computations, Prentice Hall Professional Technical Reference, 1977.