UNDERDETERMINED BLIND SEPARATION FOR SPEECH IN REAL ENVIRONMENTS WITH F0 ADAPTIVE COMB FILTERING

†‡*Federico Flego,* †*Shoko Araki,* †*Hiroshi Sawada,* †*Tomohiro Nakatani and* †*Shoji Makino*

‡flego@itc.it

†NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan

‡ITC-irst (Centro per la Ricerca Scientifica e Tecnologica) I-38050 Povo - Trento, Italy

ABSTRACT

This paper proposes a method for separating speech signals when there are more signals than sensors. The underdetermined scenario has already been investigated, exploiting the sparseness of speech signals. These methods employ binary masks to extract the signals, and therefore, the extracted signals contain loud musical noise. To mitigate this side effect, a combined continuous mask approach and post processing scheme is proposed here. First, the time-frequency points at which each source is active are estimated based on the sparseness assumption. These values are used to design continuous masks, instead of conventional binary masks, to extract target speakers. Then, using the fundamental frequency (F0) estimated from each separated speaker, F0 adaptive comb filters are tuned and used to further enhance the separation performance and sound quality of the output. Experimental results show that the proposed combination of a continuous mask and F0 adaptive comb filtering reduces the detrimental musical noise effect under both anechoic and reverberant conditions with T_R=130 ms.

1. INTRODUCTION

Blind source separation (BSS) refers to the problem of estimating original source signals from their linear mixtures. The general approach does not need a priori information about the sources or mixing process, or about the mixing matrix, sensor or speaker positions, and the only assumption is the statistical independence of the source signals [1].

Another important distinction is between experimental setups in which the number of sensors is equal to or greater than the numbers of sources, i.e. *determined* or *overdetermined* case, and situations where the source signals outnumber the sensors, i.e. *underdetermined* case.

The scheme considered here refers to the underdetermined case, where sources are speech signals and both the sensor spacing and the number of sources are known. Both anechoic and reverberant scenarios are addressed and more attention is paid to reverberant conditions. We assume that the source signals are sparse in the time-frequency domain, that is, we believe the sources rarely overlap.

A binary mask approach has been proposed for the underdetermined case [2]. However, as pointed out in [3], this method results in too much discontinuous zero-padding of the extracted signals, producing distortion and musical noise. This side effect is clearly visible when high energy regions (i.e. formants of voiced segments) that belong to different speakers overlap. As a result, the original signal structure deteriorates. However, it can be partially recovered by exploiting knowledge of speech characteristics such as the fundamental frequency (F0) information.

The system presented here works in two steps. First signals are separated by means of continuous time-frequency masks, based on a linear interpolation of direction of arrival (DOA) values. Then F0 information from each speaker is used to tune comb filters, which enhance the harmonic structure of the target signals while filtering out the residuals of interfering speakers.

2. BSS SYSTEM SETUP

A commonly used setup for a BSS system in a real environment considers M sensors observing N signals, which are modeled as convolutive mixtures $x_j(n) = \sum_{i=1}^{N} \sum_{l=1}^{L} h_{ji}(l) s_i(n-l+1)$, $j = 1, \cdots, M$. Using n to indicate a time index, $s_i(n)$ represents the i-th source, $x_j(n)$ the signal observed by the j-th sensor, and $h_{ji}(n)$ the room impulse response, which models the delay and reverberation room effects from the i-th source to the j-th sensor. Here, the underdetermined case is addressed, that is, $N > M$, with $N = 3$ and $M = 2$ and separation is carried out in the time-frequency domain. In this domain speech signals sparseness can be assumed, so that convolutive mixtures turn into instantaneous mixtures $\mathbf{X}(\omega, m) = \mathbf{H}(\omega)\mathbf{S}(\omega, m)$, where ω and m are frequency and frame indexes, respectively. $\mathbf{H}(\omega)$ is a 2×3 mixing matrix whose j, i-th component represents the transfer function from the i-th source to the j-th sensor. $\mathbf{S}(\omega, m) = [S_1(\omega, m), S_2(\omega, m), S_3(\omega, m)]^T$ and $\mathbf{X}(\omega, m) = [X_1(\omega, m), X_2(\omega, m)]^T$ denote short-time Fourier transforms of sources and observed signals, respectively. Using only information relative to observations $X_j(\omega, m)$, and knowledge of the sensor spacing, we can estimate separated signals $\mathbf{Y}(\omega, m) = [Y_1(\omega, m), Y_2(\omega, m), Y_3(\omega, m)]^T$ by means of the time-frequency masking method described in Sec. 3. Afterwards, signals $Y_k(\omega, m)$ are transformed back to the time domain providing $y_k(n)$, $k = 1, \cdots, N$, which will be used as inputs for the F0 based post-processing step described in Sec. 4. The latter approach is based on adaptive comb filtering, which exploits F0 variations of voiced segments to remove the residuals of interfering speakers and to enhance the target signal.

3. TIME-FREQUENCY MASKING METHOD

3.1. Conventional binary masks

Several methods based on source sparseness have been proposed for solving the underdetermined BSS problem [2, 4]. Sparseness implies that most of the signal samples can be considered null in a certain domain, thus making it possible to assume that sources overlap at rare intervals [5]. Given that assumption, each target speaker can be extracted by selecting just those time-frequency

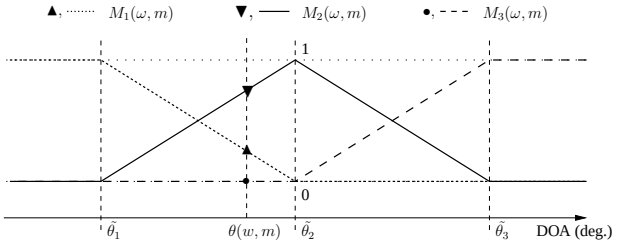

Figure 1: *Continuous mask: linear interpolation.*

bins at which the speaker is considered to be active or predominant from the mixture.
One way to localize such time-frequency bins is to compute the phase difference $\varphi(\omega,m) = \angle\frac{X_1(\omega,m)}{X_2(\omega,m)}$ between microphones observations $X_1(\omega,m)$ and $X_2(\omega,m)$. Using $\varphi(\omega,m)$, we can estimate the DOA for each time-frequency bin by computing $\theta(\omega,m) = \cos^{-1}\frac{\varphi(\omega,m)c}{\omega d}$, where c is the speed of sound and d is the microphone spacing.
For each frequency index, computing the histogram of $\theta(\omega,m)$ reveals three peaks centered approximately on the actual DOA of the sources, which can therefore be estimated by employing a clustering algorithm such as k-means. Let the centroid of each cluster be $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ where $\tilde{\theta}_1 \leq \tilde{\theta}_2 \leq \tilde{\theta}_3$.
Conventional methods based on the sparseness assumption define the binary mask as

$$M_k(\omega,m) = \begin{cases} 1 & \tilde{\theta_k} - \Delta \leq \theta(\omega,m) \leq \tilde{\theta_k} + \Delta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and extract each target signal by employing $Y_k(\omega,m) = M_k(\omega,m)X_j(\omega,m)$, j=1 or 2, $k = 1,\ldots,N$.
Δ is an extraction range parameter that determines the tradeoff between separation performance and sound fidelity.

3.2. Continuos mask

Although effective, the above binary mask approach introduces musical noise [3]. To mitigate this side effect, a continuous mask is used in place of a binary mask.
The idea is that if the DOA cannot be properly estimated for a particular time-frequency bin, the sparseness assumption is not verified for that particular value in the mixture.
The distance of each $\theta(\omega,m)$ from centroid $\tilde{\theta}_i$ is used as a "reliability" indicator for the underlying mixture value $X_j(\omega,m)$ for computing $Y_i(\omega,m)$. Consequently $M_k(\omega,m)$ will be assigned a value proportional to that distance.
Thus, each continuous mask $M_k(\omega,m)$ is now designed using linear interpolation, as shown in Fig. 1. Symbols ▲, ▼ and •, show the values assigned to masks M_1, M_2 and M_3, respectively, for the specific $\theta(\omega,m)$ that is being considered.
Other solutions to the linear interpolation are polynomial interpolation or directivity pattern based masks as described in [6].

4. POST PROCESSING

Each signal $y_k(n)$ extracted by means of continuous masks accounts for the target speaker $s_i(n)$, $i = k$, and a certain amount of residual interference due to interfering speakers $s_i(n)$, $i \neq k$. To improve separation and sound quality, thus reducing musical noise, an extra processing stage is employed as shown in Fig. 2.

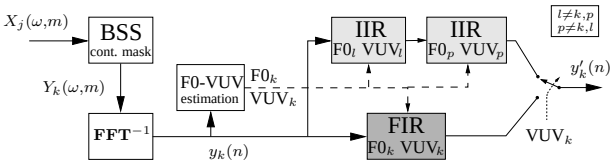

Figure 2: *F0 driven, comb filtering scheme.*

In the scheme presented here, the FO-VUV estimation block is responsible for estimating both the fundamental frequency and the voiced/unvoiced information from each of the extracted signals $y_k(n)$.
Each signal FO$_k$ will then be used to tune one different adaptive FIR or IIR filter, which will be active only on voiced segments indicated by the VUV$_k$ signal with which it is controlled.
The FIR filter is responsible for the harmonic enhancement of the target speaker $y_k(n)$, while IIR filters suppress interference caused by the other speakers in the mixture.
The final output $y'_k(t)$ is obtained by selecting the FIR filter output for speech segments labeled as voiced, and the IIR filter output for unvoiced segments. Signal VUV$_k$ drives the selection.

4.1. Harmonic enhancement of target speaker

The voiced sections of the signal under consideration, $y_k(n)$, are filtered with an adaptive FIR comb filter [7], whose impulse response $h(n)$ is shown in Fig. 3, with its coefficients a_i indicated with a filled circle.
F0 values are not constant during voiced segments of speech signals as shown in the figure, where successive pitch periods are indicated with T_{m-1}, T_m, T_{m+1} and T_{m+2}, and $\mathrm{T}_r \neq \mathrm{T}_f$, for $r \neq f$.
To take account of pitch values fluctations, the spacing between the filter impulse response values a_i, is continuously adjusted to coincide with the spacing of the individual pitch periods T_r of the waveform being processed.
The pitch period length, at each time instant, is provided by the $F0_k$ signal, which tunes the filter. The effect of this filtering procedure is that of averaging successive pitch periods of the target speaker, so that they will add constructively. Since residual components from interfering speakers do not exhibit present such periodic behaviour, they will be further reduced by the averaging procedure. This results in the restoration of harmonic components' continuity, which reduces musical noise.
We consider this filter to be more suitable for harmonic enhancement because of its fast adaption to F0 fluctuations and linear phase characteristics. Values a_i are the coefficients of a Hanning window of length N_{FIR}, and the filter frequency response is showed at the top of Fig. 4.

4.2. Removal of harmonics of interfering speakers

Unvoiced segments of the target speaker $y_k(n)$ during which competing speakers are voicing, are filtered with an adaptive IIR comb filter [8], with a transfer function given by

$$H(z) = \frac{\prod_{k=1}^{N_{\mathrm{IIR}}}(1 + \alpha_k z^{-1} + z^{-2})}{\prod_{k=1}^{N_{\mathrm{IIR}}}(1 + \rho\alpha_k z^{-1} + z^{-2})},$$

where $\alpha_k = -2\cos(k\omega_0)$, $\rho < 1$, and $\omega_0 = 2\pi F0$. N_{IIR} determines the number of harmonics to be cancelled out.

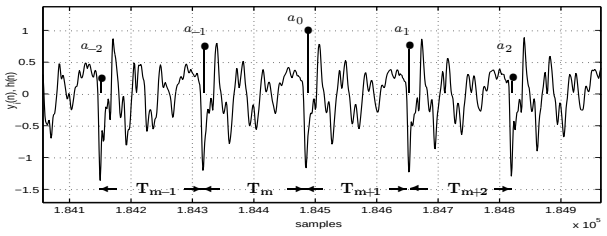

Figure 3: *Adaptive FIR filter and speech waveform with varying pitch period.*

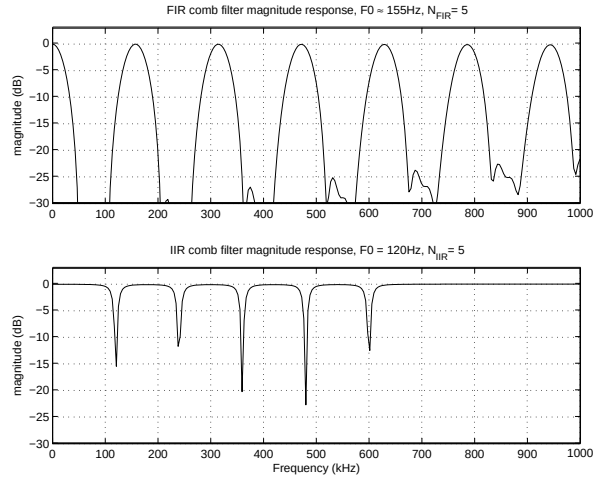

Figure 4: *Frequency response of FIR and IIR comb filters.*

The filtering is employed twice, first setting ω_0 at the F0$_l$ values of the first interfering speaker, ($l \neq k, p$), then with the F0$_p$ values of the second interfering speaker, ($p \neq k, l$). In this way, harmonics relative to the voiced segments of interfering speakers s_i, $i \neq k$, are removed from signal $y_k(n)$.

Despite its nonlinear phase characteristic, this filter is suited for harmonics removal. This because it provides a more abrupt and higher cutoff ratio in the frequency locations of interest (see bottom of Fig. 4) than its FIR counterpart. The latter in fact, must have a short impulse response to satisfy the quasi-stationarity assumption for voiced segments.

5. EXPERIMENTAL SETUP

Fig. 5 shows the setup used for the experiments. To simulate an anechoic environment, i.e. T_R = 0 ms, we used the mixing matrix $H_{ji}(\omega) = \exp(j\omega\tau_{ji})$, where $\tau_{ji} = \frac{d_j}{c}\cos\theta_i$, d_j is the positon of the j-th microphone, and θ_i is the direction of i-th source. For the reverberant case, the speech data was convolved with impulse responses recorded in a real room with a reverberation time T_R = 130 ms.

We used the Keele database [9] to test the system performance. The original audio files were downsampled to $8kHz$, delayed and added in different combinations to form 20 mixtures, each 10 seconds in length.

The sampling rate was 8 kHz while the DFT frame size and frame shift used to compute $\mathbf{X}_j(\omega, m)$, were set at 512 and 256 samples, respectively.

	Keele	Anechoic $y_k(n)$	Echoic $y_k(n)$
GER (%)	0.93	3.09	18.11
RMSE (%)	2.50	3.00	3.41

Table 1: Performance of the REPS pitch extraction algorithm, applied to the original Keele database, and to the output of a continuous mask separation system for the echoic and anechoic scenarios. GER (Gross Error Rate) is measured as the percentage of the F0 estimates which differ more than 20% respect to the actual F0 values. The root-mean-squared error (RMSE) or "fine pitch error", is computed on the remaining F0 estimates.

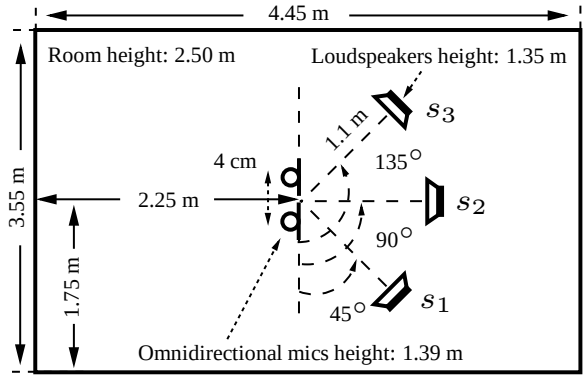

Figure 5: *Room for reverberant tests.*

Both F0 and VUV data are estimated using the *Ripple-Enhanced Power-Spectral-based* (REPS) algorithm [10]. This algorithm was demonstrated to be very effective in pitch estimation, even for reverberant signals corrupted by musical noise such as those we are dealing with. F0 values are used to tune the FIR and IIR comb filters as described in Sec. 4.

For F0 estimation the frame size and frame shift were set at 336 and 8 samples, respectively. The filter parameters were tuned to achieve the best results: $N_{\mathrm{FIR}} = 5$, $N_{\mathrm{IIR}} = 5$ and $\rho = 0.995$.

In the binary mask approach, Δ was set so that all the values belonging to each estimated cluster, were used in the design of the corresponding mask. This also implies that the assignment of each mixture bin is mutually exclusive, that is, every bin from the mixture is used for only one target speaker reconstruction.

6. PERFORMANCE EVALUATION CRITERIA

Signal to interference ratio (SIR) and signal to distortion ratio (SDR) were chosen as measures of separation performance and sound quality, respectively:

$$\mathrm{SIR}_k = 10\log\frac{\sum_n y^2_{ks_k}(n)}{\sum_n(\sum_{k\neq i} y_{ks_i}(n))^2}, \quad (2)$$

$$\mathrm{SDR}_k = 10\log\frac{\sum_n x^2_{js_k}(n)}{\sum_n(x_{js_k}(n) - \alpha y_{ks_k}(n-D))^2}, \quad (3)$$

where y_k is the estimation of s_k, and y_{ks_i} is the k-th separating system output when only s_i is active and s_l, $l \neq i$ is silent; x_{js_k} is the observation provided by microphone j when only s_k is active. α and D are parameters to compensate for the amplitude and phase difference between x_{js_k} and y_{ks_k}. To evaluate the

	SIR (dB)	SDR (dB)	SIR (%)	SDR (%)
	Binary mask			
Anechoic	13.50	11.46		
Echoic	10.65	8.92		
	Continuous mask		Rel. improvement %	
Anechoic	13.86	12.06	2.67	5.24
Echoic	10.55	9.83	-0.93	10.20
	Continuous mask + PP		Rel. improvement %	
Anechoic	14.55	11.84	7.78	3.32
Echoic	11.38	9.50	6.85	6.50

Table 2: BSS results using continuous masks, binary masks, and a combination of continuous mask and F0 based post processing (PP).

performance of the proposed methods, SIR and SDR are computed using measurements from both microphones and the best value is retained.

7. EXPERIMENTAL RESULTS

7.1. Binary mask

For comparison purposes, the binary mask approach is assumed here as the baseline system, with the results shown at the top of Table 2. In the anechoic scenario, signals better satisfy the sparseness assumption, providing histograms of $\theta(\omega, m)$ with well localized and sharp peaks along the θ axes, making the estimation of $\tilde{\theta}_i$ in (1) more reliable.

By contrast, in the echoic scenario, reverberaration causes signals to overlap more in the time-frequency domain. This makes the estimation of $\theta(\omega, m)$ more difficult and less reliable. This in turn explains the performance degradation shown in the table.

7.2. Continuous mask

When the estimated DOA for a particular mixture time-frequency bin, differs considerably from any estimated centroid $\tilde{\theta}_i$, the probability of speaker superposition is considered to be higher than when the DOA coincides with one of the centroids. In such a case, this time-frequency bin will generate distortion in the speaker signal selected for the target, whereas there will be information missing from the spectrograms of the other extracted signals.

To partially overcome the latter problem, continuous masks are employed, assigning a weight for each mask that is linearly proportional to the distance of the estimated DOA from each centroid for the bin under consideration.

The results we obtained with this approach are shown in the centre of Table 2, and demonstrate the advantage of using continuous masks, particularly in the echoic case where DOA estimation is more difficult. The table also indicates that signal distorsion (SDR) is generally reduced while SIR shows little overall improvement.

A Gaussian interpolation in the mask design was also tested with similar results.

7.3. Comb filtering + post processing

We applied comb filtering to the output obtained with continuous masks and this provided the results shown at the bottom of Table 2. While the use of the continuous masks principally improved the SDR, the main achievement of the comb filtering was to increase the SIR at the expense of the SDR. This proves the effectiveness of the comb filtering scheme for eliminating interference and restoring signal harmonics, in both anechoic and reverberant scenarios.

8. CONCLUSION

A combined continuous mask approach and adaptive comb filtering scheme was proposed for BSS when speech signals outnumber sensors. This method proved to be effective for mitigating the adverse effect of loud musical noise induced by binary mask based BSS systems. Both the anechoic and reverberant scenarios ($T_R = 130$ ms) were tested. We obtained improvements in both the SIR and SDR ratios, and kept the overall computational complexity low.

9. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, 2001.

[2] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in *Proc. ICASSP 2002*, 2002, vol. 1, pp. 529–532.

[3] S. Araki, S. Makino, A. Blin, Mukai R., and H. Sawada, "Undetermined blind separation for speech in real environments with sparseness and ica," in *Proc. ICASSP 2004*, 2004, vol. 3, pp. 881–884.

[4] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time fourier transform," in *Proc. ICA 2000*, 2000, pp. 87–92.

[5] A. Blin, S. Araki, and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix combination," in *Proc. IWAENC 2003*, 2003, pp. 211–214.

[6] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica," in *Proc. ICA 2004*, 2004, pp. 898–905.

[7] R. H. Frazier, S. Samsam, Braida L. D., and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. ICASSP 1976*, 1976, pp. 251–253.

[8] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1124–1138, Oct. 1986.

[9] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. EuroSpeech 1995*, 1995, pp. 837–840.

[10] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *Journal of the Acoustical Society of America (JASA)*, vol. 116, no. 6, pp. 3690–3700, Dec. 2004.