Time-Frequency-Processing for ICA-Supported Speech Recognition in Multitalker Conditions

Eugen Hoffmann[1], Dorothea Kolossa[2], and Reinhold Orglmeister[1]

[1] *Technische Universität Berlin, Electronics and Medical Signal Processing Group, Einsteinufer 17, 10587 Berlin, Germany*

[2] *Ruhr-Universität Bochum, Institute of Communication Acoustics, Universitätsstr. 150, 44801 Bochum, Germany*

Zusammenfassung

Blind source separation for mixtures of acoustic signals is usually performed in the frequency domain, where source separation by independent component analysis (ICA) is applied separately in each frequency bin. However, there arise two problems. The first is obtaining a consistent ordering of the recovered signals, also known as the permutation problem. The other problem is the remaining interference in the separated signals and it can be ameliorated by applying time-frequency masking.

In this paper, we present new developments for these two main obstacles: a permutation correction based on the correntropy is described. A time-frequency mask is then introduced based on an approximation of source and noise dominance and on a consequential extension of the Wiener-filter to multichannel data. The suggested time-frequency mask leads to appreciable improvements in automatic speech recognition (ASR) performance, and other than with many standard time-frequency masks, the ASR improvements do not depend on the use of missing data speech recognition, but are achievable to almost their full extent without modifications to the decoder.

Independent Component Analysis

Acoustic signal mixtures in reverberant environments can be described by convolution of a source signal with an unknown filter matrix $\mathbf{A}$. Transformation to the frequency domain reduces the convolutions to multiplications

$$\mathbf{X}(\Omega,\tau) \approx \mathbf{A}(\Omega)\mathbf{S}(\Omega,\tau),$$

where Ω is the angular frequency, τ denotes the frame index, $\mathbf{A}(\Omega)$ is the mixing system in the frequency domain, $[S_1(\Omega,\tau),\ldots,S_N(\Omega,\tau)]$ represents the source signal, and $[X_1(\Omega,\tau),\ldots,\ X_N(\Omega,\tau)]$ the observed signals. Then, for each frequency bin, only an instantaneous ICA problem needs to be solved. For this purpose, the FastICA algorithm has been used [1].

Permutation Correction

The filter matrices calculated by ICA can be randomly permuted. To sort the permuted signals in each frequency bin, we propose a new method based on the correntropy definition in [2]. To solve the permutation problem, for each frequency bin k, the non-permuted signal $\mathbf{Y}$,

$$\mathbf{Y}(\Omega_k,\tau) = \arg\max_{r=1\ldots R} V\left(\left|\mathbf{Y}^r(\Omega_k,\tau)\right|, \left|\hat{\mathbf{Y}}(\Omega,\tau)\right|\right)$$

has to be found, where $\mathbf{Y}^r(\Omega_k,\tau)$ is the r-th among R possible permutated signal vectors, $\hat{\mathbf{Y}}(\Omega,\tau)$ is an average of L already corrected bins $\frac{1}{\hat{L}-k}\sum_{l=k+1}^{\hat{L}}|\mathbf{Y}(\Omega_l,\tau)|$, with $\hat{L}=\min(k+1+L, N_{FFT}/2+1)$ and $V(\cdot,\cdot)$ is the correntropy

$$V\left(\left|\mathbf{Y}^r(\Omega_k,\tau)\right|, \left|\hat{\mathbf{Y}}(\Omega,\tau)\right|\right) = E_\tau\left[\kappa\left(\left|\mathbf{Y}^r(\Omega_k,\tau)\right|, \left|\hat{\mathbf{Y}}(\Omega,\tau)\right|\right)\right]$$

with the kernel-function $\kappa(\cdot,\cdot)$ and

$$E_\tau\left[\kappa\left(\left|\mathbf{Y}^r(\Omega_k,\tau)\right|, \left|\hat{\mathbf{Y}}(\Omega,\tau)\right|\right)\right] \approx \frac{1}{T}\sum_{\tau=1}^{T}\kappa\left(\left|\mathbf{Y}^r(\Omega_k,\tau)\right|, \left|\hat{\mathbf{Y}}(\Omega,\tau)\right|\right)$$

where T is the number of frames. In the following, the Laplace kernel was chosen for $\kappa(\cdot,\cdot)$ [3].

Postprocessing

In this section, a time-frequency (TF) mask is described that minimizes the remaining noise components. The mask is based on target

$$\lambda_{S_i}(\Omega,\tau) = \frac{\left\|\mathbf{a}_i^H(\Omega)Y_i(\Omega,\tau)\right\|^2}{\sum_{k=1}^{N}\left\|\mathbf{a}_k(\Omega)Y_k(\Omega,\tau)\right\|^2}$$

and noise dominances

$$\lambda_{D_i}(\Omega,\tau) = 1 - S_i(\Omega,\tau),$$

which are calculated based on the power ratio defined in [4]. $\mathbf{a}_i(\Omega)$ denotes the i-th column of the estimated mixing matrix and $Y_i(\Omega,\tau)$ is the i-th unmixed signal.

These two measures are used to approximate the a priori signal-to-noise ratio (SNR)

$$\xi_i(\Omega,\tau) \approx \left(\frac{\lambda_{S_i}(\Omega,\tau)}{a\lambda_{D_i}(\Omega,\tau)}\right)^p + b,$$

where a, b and p are the approximation parameters. In the next step, the a priori SNR is used for calculating the Wiener filter gain for noise suppression

$$M_i(\Omega,\tau) = \frac{\xi_i(\Omega,\tau)}{1+\xi_i(\Omega,\tau)}. \quad (1)$$

The enhanced signal $\tilde{Y}_i(\Omega,\tau)$ is finally obtained by

$$\tilde{Y}_i(\Omega,\tau) = M_i(\Omega,\tau)Y_i(\Omega,\tau).$$

Robust Speech Recognition

Automatic speech recognition can suffer from TF masking, but when an error variance, or *uncertainty* is estimated along with the signal and used in decoding, as described e.g. in [5], performance often increases. An overall block diagram is given in Fig. 1.

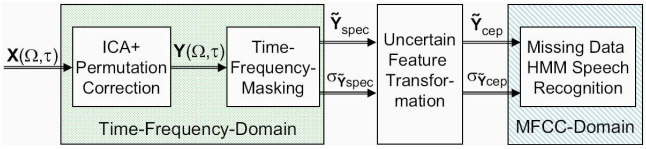

Abbildung 1: Block diagram with Data Flow.

In this paper, we have estimated uncertainties $\sigma(\Omega,\tau)$ by

$$\sigma_{\tilde{Y}\,spec}(\Omega,\tau) = \alpha \cdot |\hat{S}^2(\Omega,\tau) - \tilde{Y}^2(\Omega,\tau)|; \qquad (2)$$

with α as an uncertainty weight, $\tilde{Y}(\Omega,\tau)$ as the masked ICA output and $\hat{S}(\Omega,\tau)$ as the source signal reference, which is calculated using a more aggressive mask of the same type.

Experiments and Results

To evaluate the proposed approaches, different real room recordings of audio files from the TIDigits database [6] were made in a mildly reverberant ($T_R \approx 160$ ms) and noisy lab room. The distances L_i between loudspeakers and microphones were varied between 0.9 and 3 m. The experimental conditions are summarized in Table 1, where N indicates the number both of speakers and microphones. The algorithms were tested on the room recordings, which were first transformed to the frequency domain at a resolution of $N_{\mathrm{FFT}} = 1024$. For calculating the STFT, the signals were divided into overlapping frames using a Hanning window with an overlap of $3/4 \cdot N_{\mathrm{FFT}}$. The parameter settings for the evaluated TF masks are different for the use in noise suppression, where they are set to $a = 0.2, b = 0.1$, and $p = 0.5$, and for the use as the reference $\hat{S}$ in Eq. (2), in which case they are chosen as $a = 40, b = 0.01$, and $p = 2$. α was set to 0.6.

Tabelle 1: Mixture recording parameters: Number of sources N, distances L_i between loudspeakers and array center, directions of arrival θ_i relative to broadside.

	Mix. 1	Mix. 2	Mix. 3	Mix. 4	Mix. 5
N	2	3	2	2	3
$[L_i]$	[2.0, 2.0]	0.9 each	[1.0, 3.0]	[0.9, 0.9]	0.9 each
$[\theta_i]$	[75°, 165°]	[30°, 80°, 135°]	[50°, 100°]	[50°, 115°]	[40°, 60°, 105°]

Performance Measurement

The signal-to-interference ratio (SIR) was used as a measure of the separation performance and the signal-to-distortion ratio (SDR) as a measure of signal quality [7]. To evaluate ASR performance, the number of word reference labels (W), of substitutions (S), insertions (I) and deletions (D) are counted. From these values, two performance figures are obtained, the *correctness* $PC = \frac{W-D-S}{W}$ and the *accuracy* $PA = \frac{W-D-S-I}{W}$. Correctness indicates distortions of the desired speech signal, but as it ignores insertion errors, it does not penalize audibility of the interfering speaker. Therefore, the accuracy is also needed to measure how well interferences are suppressed from the recognition output.

The results of ICA with and without TF masking are shown in Tab. 2. All mixtures were separated with the FastICA algorithm and subsequently the suggested TF mask was applied. As can be seen, the postmask achieves an average SDR improvement of almost 2dB, while incurring an SIR loss of only 0.2dB. This good ratio between added interference suppression and loss in signal quality has a positive influence on the ASR results shown in Tab 3. As it can be seen, using no uncertainties (NU), the suggested TF mask increases the recognition accuracy by an average of 6% absolute. When the observation uncertainties are also applied, using uncertainty decoding (UD) or modified imputation (MI) for more robust ASR [5], an absolute 3% are gained additionally.

Tabelle 2: Mean value of output ΔSIR/SDR in dB.

ICA, no Mask	ICA & mask	ICA & mask for uncertainty estimation
6.2/5.5	8.1/5.3	17.2/1.7

Tabelle 3: Average recognition rate over all speakers

	Correctness (PC)			Accuracy (PA)		
Mask	NU	UD	MI	NU	UD	MI
off	0.763	0.775	0.771	0.731	0.745	0.741
on	0.842	0.850	0.856	0.794	0.816	0.826

Conclusions

A new post-mask for ICA has been suggested to improve recognition performance in reverberant multi-source conditions. This post-mask increases ASR performance even without modifications to the decoder, but best performance, with an overall word error rate reduction of 35%, is achieved when the mask is applied in conjunction with uncertainty-of-observation techniques.

Literatur

[1] A. Hyvärinen and E. Oja. "A Fast Fixed-Point Algorithm for Independent Component Analysis". in *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[2] I. Santamaria, and P. P. Pokharel, and J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization." *IEEE Trans. Signal Proc.*, 54(6), pp. 2187—2197, 2006.

[3] B. Schölkopf, and A. J. Smola, " Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond." MIT Press, 2001.

[4] H. Sawada, S. Araki, and S. Makino, "Measuring Dependence of Bin-wise Separated Signals for Permutation Alignment in Frequency-domain BSS," *Proc. ISCAS*, pp. 3247–3250, May 2007.

[5] D. Kolossa, S. Araki, M. Delcroix, T. Nakatani, R. Orglmeister, and S. Makino, "Missing Feature Speech Recognition in a Meeting Situation with Maximum," *Proc. ISCAS*,pp. 3218–3221, April 2008.

[6] R. G. Leonard, "A Database for Speaker-Independent Digit Recognition", *Proc. ICASSP*, Vol. 3, p. 42.11, 1984.

[7] H. Sawada, S. Araki, R. Mukai and S. Makino, "Blind Extraction of a Dominant Source from Mixtures of Many Sources using ICA and Time-Frequency Masking," *Proc. ISCAS*, pp. 5882–5885, May 2005.