Missing Feature Speech Recognition in a Meeting Situation with Maximum SNR Beamforming

Dorothea Kolossa*†, Shoko Araki†, Marc Delcroix†, Tomohiro Nakatani†, Reinhold Orglmeister*, Shoji Makino†
*Electronics and Medical Signal Processing, TU Berlin
10587 Berlin, Germany
Email: d.kolossa@ee.tu-berlin.de
†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: {shoko,marc.delcroix,nak,maki}@cslab.kecl.ntt.co.jp

***Abstract*— Especially for tasks like automatic meeting transcription, it would be useful to automatically recognize speech also while multiple speakers are talking simultaneously. For this purpose, speech separation can be performed, for example by using maximum SNR beamforming. However, even when good interferer suppression is attained, the interfering speech will still be recognizable during those intervals, where the target speaker is silent. In order to avoid the consequential insertion errors, a new soft masking scheme is proposed, which works in the time domain by inducing a large damping on those temporal periods, where the observed direction of arrival does not correspond to that of the target speaker. Even though the masking scheme is aggressive, by means of missing feature recognition the recognition accuracy can be improved significantly, with relative error reductions in the order of 60% compared to maximum SNR beamforming alone, and it is successful also for three simultaneously active speakers. Results are reported based on the SOLON speech recognizer, NTT's large vocabulary system [1], which is applied here for the recognition of artificially mixed data using real-room impulse responses and the entire clean test set of the Aurora 2 database.**

I. Introduction

Many source separation methods are optimized for situations, in which the number of speakers is fixed, and known beforehand. However, this assumption is unrealistic for many potential real-world applications, such as automatic meeting transcriptions. For such situations, in which the number and position of active speakers often changes over time, more flexible source separation architectures are needed. In this paper, the use of maximum SNR beamforming [2] with adaptive online clustering is investigated for application together with automatic speech recognition and it is modified to also include a time-masking which improves separation performance.

However, such quickly time-varying nonlinear masking is generally detrimental for speech recognition, therefore, instead of using a standard HMM speech recognition engine, the SOLON large vocabulary speech recognizer has been modified to be able to include additional uncertainty information in the recognition process. This information is won from the preprocessing module, and passed on to the recognizer, as shown in the overview in Figure 1. Since the recognition takes places on mel-cepstrum features $\tilde{Y}(c,\tau)$, rather than the STFT-coefficients $\tilde{y}(f,\tau)$ which are computed by the preprocessing stage, an additional module for transforming the features $\tilde{y}(f,\tau)$ with associated uncertainties $\sigma_{\tilde{y}}(f,\tau)$ from the spectrum to the mel-cepstrum features $\tilde{Y}(c,\tau)$ with their uncertainties $\sigma_{\tilde{Y}}(c,\tau)$ is necessary. This transformation is achieved here by means of the unscented transform.

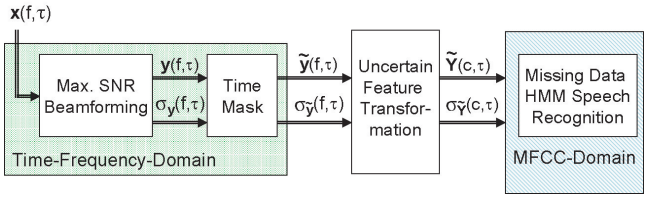

Fig. 1. Overall Structure of the System

II. Mixing Model

It is assumed here, that noisy, convolutive mixing of N speech sources takes place. This means that the M observed microphone signals $x_j(t)$, $j = 1 \dots M$ are composed of

$$x_j(t) = \sum_{k=1}^{N} \sum_{l} h_{jk}(l) s_k(t-l) + n_j(t), \qquad (1)$$

where h_{jk} stands for the room impulse response from source k to sensor j, and where n_j is the sensor noise at sensor j. After a T-point short time Fourier transform, the mixing model (1) simplifies to

$$x_j(f,\tau) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau) + n_j(f,\tau), \qquad (2)$$

where $f \in \left\{0, \dots, \frac{T-1}{T} f_s\right\}$, f_s is the sampling frequency and τ denotes the frame index.

III. Maximum SNR Beamforming

Maximum SNR beamforming attempts to maximize the Signal-to-Noise-Ratio (SNR) at the beamformer output. In order to achieve this, it is necessary first, to distinguish target-active periods $\mathcal{P}_T^k$ from target-inactive-periods $\mathcal{P}_I^k$ for each source k. As described in [2], this can be achieved by a three-stage process.

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

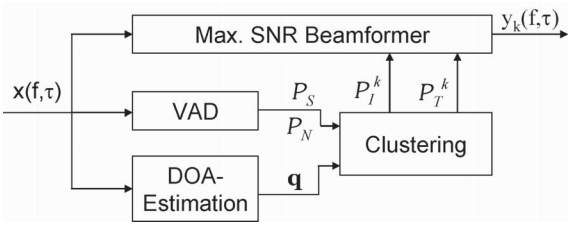

Fig. 2. Block diagram of maximum SNR beamformer

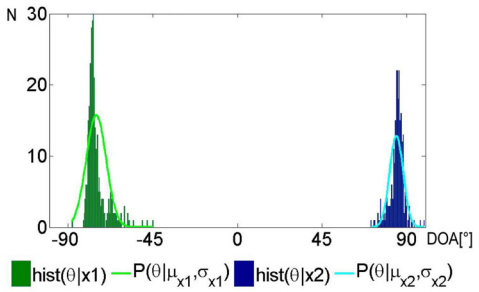

Fig. 3. Histogram of estimated DOAs and approximation with Gaussian distribution.

- First, noise only periods are determined by statistical, model-based voice activity detection according to [3]. This results in a distinction between speech frames $\mathcal{P}_S$ and noise frames $\mathcal{P}_N$.
- Secondly, the generalized cross correlation method with phase transform (GCC-PHAT) [4] is applied for determining time-differences of arrival (TDOAs) $\mathbf{q}$ in each of the frames.
- Finally, all speech frames $\mathcal{P}_S$ are clustered according to their TDOA by means of an online clustering method, which does not require advance knowledge about the number of active sources [5].

This final clustering yields the target-active periods $\mathcal{P}_T^k$ for each of the sources k. Interference periods are assumed for each source, whenever the source is not deemed active, thus $\mathcal{P}_I^k = \mathcal{P} - \mathcal{P}_T^k$ and the set of all frames $\mathcal{P}$ is therefore partitioned into two disjoint sets for each of the sources. Using target-active-periods $\mathcal{P}_T^k$ and source-inactive periods $\mathcal{P}_I^k$ for source k, the beamformer weights are obtained by optimizing the ratio between the output energy in target-active periods and the energy in interference periods for each of the sources $k = 1 \dots N$ according to

$$\begin{aligned}\mathbf{w}_{k,opt}(f) &= \arg\max_{\mathbf{w}_k(f)} \frac{\mathrm{E}(|y_k(f,\tau)|^2)_{\mathcal{P}_T^k}}{\mathrm{E}(|y_k(f,\tau)|^2)_{\mathcal{P}_I^k}} \\ &= \arg\max_{\mathbf{w}_k(f)} \frac{\mathbf{w}_k^H(f)\mathbf{R}_T^k(f)\mathbf{w}_k(f)}{\mathbf{w}_k^H(f)\mathbf{R}_I^k(f)\mathbf{w}_k(f)}. \end{aligned} \quad (3)$$

The correlation matrix for target-active periods $\mathbf{R}_T^k$ is calculated by

$$\begin{aligned}\mathbf{R}_T^k &= \mathrm{E}\left(\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^H\right)_{\mathcal{P}_T^k} \\ &= \frac{1}{|\mathcal{P}_T^k|}\sum_{\tau\in\mathcal{P}_T^k}\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^H \end{aligned} \quad (4)$$

and $\mathbf{R}_I^k$ is obtained in the same manner via

$$\begin{aligned}\mathbf{R}_I^k &= \mathrm{E}\left(\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^H\right)_{\mathcal{P}_I^k} \\ &= \frac{1}{|\mathcal{P}_I^k|}\sum_{\tau\in\mathcal{P}_I^k}\mathbf{x}(f,\tau)\mathbf{x}(f,\tau)^H. \end{aligned} \quad (5)$$

Finally, the optimum beamformer weights $\mathbf{w}_{k,opt}(f)$ are obtained by differentiating (3) with respect to $\mathbf{w}_k$. This results in the eigenvector problem

$$\mathbf{R}_T^k(f)\mathbf{w}_k(f) = \lambda(f)\mathbf{R}_I^k(f)\mathbf{w}_k(f), \quad (6)$$

where $\mathbf{w}_{k,opt}(f)$ is the eigenvector associated with the largest eigenvalue $\lambda_{max}(f)$. This generalized eigenvector problem can also be simplified to an eigenvalue problem by left-multiplying both sides with $\mathbf{R}_I^k(f)^{-1}$. Finally, the output signal for source k is obtained by

$$y_k(f,\tau) = \mathbf{w}_{k,opt}(f)^H\mathbf{x}(f,\tau). \quad (7)$$

Here, the signal vector $\mathbf{x}(f,\tau)$ is composed of all microphone signals $[x_1(f,\tau)\dots x_M(f,\tau)]^T$.

IV. Time Masking

Maximum SNR beamforming by itself already leads to a significant suppression of the interfering speaker, for example, in [2], a signal to interference ratio (SIR) of 12dB was attained for a reverberation time of 350ms. However, for the task of robust speech recognition, greater SIR values are needed in order to avoid insertion errors during those periods, where the target speaker is inactive. Therefore, in order to improve recognition results, a time-varying damping was applied to the beamformer output, which introduces additional damping according to

$$\tilde{y}_k(f,\tau) = D_k(\tau)\cdot y_k(f,\tau). \quad (8)$$

The damping factor D_k is obtained by comparing the direction of arrival (DOA) for each time frame, $\theta(\tau)$, with a source DOA model $P(\theta(\tau)|s_k)$ for source s_k. For this purpose, the TDOA which is already available for each time frame τ, is converted to a DOA value with the help of the known array geometry [6]. Then, a statistical model of source DOAs can be learned for each source k by fitting a model to the observed DOAs in those frames where source k is active, i.e. in which $\tau \in \mathcal{P}_T^k$. As Figure 3 shows, even a simple Gaussian model may be sufficient. Such a model can be learned online, and it gives the probability $P(\theta(\tau)|s_k)$ for each frame τ and each source k. This probability was first used, to obtain a mask according to $D1_k(\tau) = P(\theta(\tau)|s_k)$. However, the probability is very quickly time-varying, so that low-energy frames within words might be corrupted by application of the mask. Therefore, a simple smoothing filter was applied to the intermediate mask $D1$ by

$$D2_k(\tau) = \frac{1}{2L+1}\sum_{i=-L}^{L} D1_k(\tau - i), \quad (9)$$

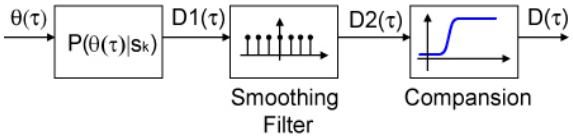

Fig. 4. Calculation of Mask from Source Probability

with $L = 7$ and a final compansion

$$D_k(\tau) = \mathrm{L}(D2_k(\tau), \alpha, \vartheta) \tag{10}$$

with the logistic function

$$\mathrm{L}(D, \alpha, \vartheta) = \alpha + \frac{1-\alpha}{1+\exp(10(\vartheta - D))} \tag{11}$$

ensures that the mask $D_k(\tau)$ remains between a small constant $\alpha > 0$, and 1 as an upper limit. The lower bound α and the threshold ϑ of L were set to $\alpha = 0.03$ and $\vartheta = 0.5$.

V. Missing Feature Recognition

A. Uncertainty Calculation

To estimate uncertainty values, it was assumed that the maximum SNR beamformer leaves residual interference in the signal, which is damped with a damping factor d. Thus, the remaining uncertainties are estimated by taking a summation over all estimated non-target outputs, each damped with the factor d, according to

$$\sigma_{\tilde{y}_k}(f,\tau) = \sum_{i=1, i\neq k}^{N} d|\tilde{y}_i(f,\tau)|. \tag{12}$$

The damping factor was set to $d = 0.1$ for all experiments.

B. Uncertainty Transformation

After the preprocessing stage, processed features $\tilde{y}_k(f,\tau)$ and associated uncertainty values $\sigma_{\tilde{y}_k}(f,\tau)$ are available in the time-frequency domain. However, since speech recognition can be performed far more robustly in other domains, using features such as the mel-cepstrum or RASTA-PLPs, a transformation of these uncertain features to the domain of speech recognition is necessary. This transformation can be calculated analytically for specific transformations, or it can be obtained in a general fashion by means of the unscented transform [7], which is the way chosen for all subsequent experiments. Then, after the unscented transform, features $\tilde{Y}_k(c,\tau)$ and associated uncertainty values $\sigma_{\tilde{Y}_k}(c,\tau)$ are available in the mel-cepstrum domain. This is the basis for missing-feature speech recognition.

C. Uncertain Recognition

Speech recognition was performed with the SOLON recognizer, NTT's large vocabulary recognition system. Two missing feature approaches, modified imputation [8] and HMM variance compensation [9], have been implemented for SOLON and were used in the tests. Both methods are developed for HMM based systems, where recognition takes place by finding the optimum HMM state sequence $[q_1, \ldots, q_E]$, which gives the best match to the feature vector sequence $[\mathbf{Y}(1), \ldots, \mathbf{Y}(E)]$ when each HMM state has an associated output probability distribution $\mathrm{p}(\mathbf{Y}|q)$. $\mathbf{Y}(i)$ denotes the i^{th} feature vector defined as $\mathbf{Y}(i) = [Y(1,i), ..., Y(F,i)]^T$, where $i = 1, ..., E$ and F and E are the number of features and the number of frames, respectively.

1) HMM Variance Compensation: In HMM variance compensation, the computation of state output probabilities is modified to incorporate frame-by-frame and feature by feature uncertainties [9]. The formulation

$$p(\tilde{\mathbf{Y}}(\tau)|q) = \int_{-\infty}^{\infty} \mathrm{p}(\tilde{\mathbf{Y}}(\tau)|\mathbf{S}(\tau))\mathrm{p}(\mathbf{S}(\tau)|q)\mathrm{d}\mathbf{S}(\tau) \tag{13}$$

leads to

$$p(\tilde{\mathbf{Y}}(\tau)|q) = N(\tilde{\mathbf{Y}}(\tau); \mu_{\mathbf{q}}, {\sigma_{\mathbf{q}}}^2 + \sigma_{\tilde{\mathbf{Y}}}(\tau)^2), \tag{14}$$

where q denotes the HMM state, with the parameters $\mu_{\mathbf{q}}$ and $\sigma_{\mathbf{q}}$, $\tilde{\mathbf{Y}}(\tau)$ is the feature vector at frame τ with the associated feature uncertainty $\sigma_{\tilde{\mathbf{Y}}}(\tau)$, and $\mathbf{S}$ stands for the clean speech feature vector.

2) Modified Imputation: In modified imputation, the idea is replacing the imputation equation, originally proposed for completely missing features in [10], by an alternative formulation, which also allows for real-valued degrees of uncertainty. Thus, whereas missing parts of feature vectors are replaced by the corresponding components of the HMM model mean $\mu_{\mathbf{q}}$ in classical imputation, modified imputation finds the maximum likelihood estimate

$$\hat{\mathbf{Y}}(\tau) \quad = \quad \arg\max_{\mathbf{Y}(\tau)} \mathrm{p}(\mathbf{Y}(\tau)|q, \mathbf{x}(\tau)). \tag{15}$$

For a single Gaussian, and assuming a flat prior for $\mathbf{Y}(\tau)$, Equation (15) leads to

$$\hat{\mathbf{Y}}(\tau) = \left(\sigma_{\tilde{\mathbf{Y}}}(\tau)^{-1} + {\sigma_{\mathbf{q}}}^{-1}\right)^{-1} \left(\mu_{\mathbf{q}}{\sigma_{\mathbf{q}}}^{-1} + \tilde{\mathbf{Y}}(\tau)\sigma_{\tilde{\mathbf{Y}}}(\tau)^{-1}\right),$$

as shown in [8], but better results are achieved with an additional uncertainty weight w according to

$$\hat{\mathbf{Y}}(\tau) = \left((w\sigma_{\tilde{\mathbf{Y}}}(\tau))^{-1} + {\sigma_{\mathbf{q}}}^{-1}\right)^{-1} \left(\mu_{\mathbf{q}}{\sigma_{\mathbf{q}}}^{-1} + \tilde{\mathbf{Y}}(\tau)(w\sigma_{\tilde{\mathbf{Y}}}(\tau))^{-1}\right)$$

and with an empirically chosen $w = 0.1$.

VI. Experiments and Results

A. Simulations with Recorded Impulse Responses

Recordings of impulse responses, made in a meeting room with the reverberation time $t_{rev} \approx$ 100ms and the layout shown in Figure 5, have been used to generate artificial mixtures of the clean test set of the Aurora 2 database. The clean test set consists of 1081 utterances with 12147 words.

B. Recognizer Parameters

The sampling rate of the data is 8kHz. Before feature extraction, the signal is highpass filtered with a cutoff frequency of 64Hz and a first order FIR preemphasis filter with the coefficient a_0=0.97 is applied. The STFT windowsize is 30ms with 20ms overlap. The acoustic model consists of speaker independent, word based HMMs trained on clean speech and

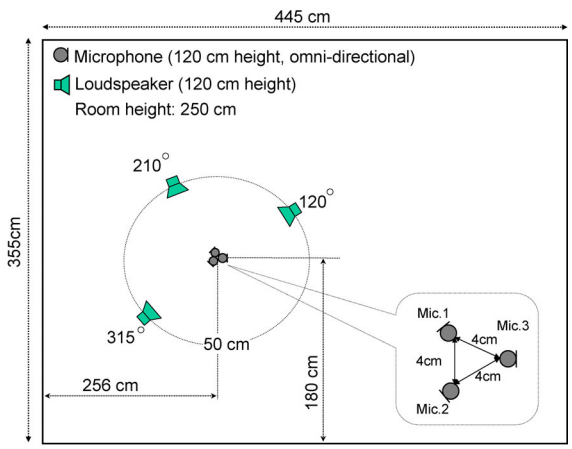

Fig. 5. Recording setup for obtaining room impulse responses

TABLE I

RESULTS FOR RECOGNITION OF AURORA TESTSET, USING MIXTURES OF TWO SPEAKERS AT 210° AND 315° IN FIGURE 5.

	S	*D*	*I*	Word Accuracy
Clean	0.9	0.6	1.1	97.4
Mixtures	20.9	2.3	48.2	28.5
MaxSNR	5.2	0.8	39.4	54.4
MaxSNR + Mask	8.0	2.2	27.5	62.3
MaxSNR + Mask + MI	6.9	3.0	8.5	81.6
MaxSNR + Mask + VC	6.9	2.9	9.7	80.5

each state uses a single Gaussian distribution to model the 39-dimensional feature vector consisting of 13 mel frequency cepstral coefficients, including the zero'th coefficient, deltas and acceleration features.

C. Recognition Results

Table I shows the results for clean data, artificial mixtures and the results of the various intermediate stages of processing. *S* stands for the percentage of substitution errors, *D* and *I* are the deletions respectively insertions and the *Word Accuracy* is obtained as $100-S-D-I$. Using maximum SNR beamforming alone increases the word accuracy from 28.5%, measured with the artificial mixtures of two speakers, to 54.4%. Most of the remaining errors are due to insertions, which, at 39.4%, make up 86% of the total error rate. Subsequent application of the masking function reduces the insertion error rate to 27.5%. Further use of the available uncertainty information leads to significant further improvements. By means of modified imputation (MI), the final word accuracy is 81.6%, whereas variance compensation (VC) leads to 80.5% word accuracy. The situation is similar in the second test setup, where artificial mixtures of three speakers are used. The results for this setup are shown in Table II. Here, again, insertion errors make up the bulk of the error rate, with 71.5% insertion errors for the mixtures and 60.4% for the beamformer output, and this main cause of recognition errors can be significantly reduced by means of time-masking. The lowest rate of errors is achieved

TABLE II

RESULTS FOR RECOGNITION OF AURORA TESTSET, USING MIXTURES OF THREE SPEAKERS.

	S	*D*	*I*	Word Accuracy
Clean	0.9	0.6	1.1	97.4
Mixtures	33.9	3.6	71.5	-9.1
MaxSNR	14.2	2.0	60.4	23.5
MaxSNR + Mask	15.5	2.9	48.8	33.2
MaxSNR + Mask + MI	14.7	4.7	28.2	52.4
MaxSNR + Mask + VC	14.4	4.4	28.7	52.5

by means of subsequent uncertain recognition, where the results of modified imputation and of variance compensation are similar, with an overall accuracy of 52.4% respectively 52.5%.

VII. CONCLUSIONS

Maximum SNR beamforming is a promising approach for the separation of speech in meeting situations, since, unlike for classical ICA, neither the number nor the location of the speakers is required to be known or constant over time. In order to use this method for meeting transcription, however, greater interference suppression is necessary especially during target inactive periods. For this purpose, a time-varying damping has shown good results, for two as well as for three mixed speech signals. Further improvements can be obtained, and especially the number of insertions can be reduced significantly, when this postprocessing is used in conjunction with missing feature recognition, which has been applied here in the decoding stage of the SOLON large vocabulary speech recognizer.

REFERENCES

[1] T. Hori, "NTT speech recognizer with outlook on the next generation: SOLON," in *Proc. NTT Workshop on Communication Scene Analysis*, 2004, pp. SP–6. [Online]. Available: www.kecl.ntt.co.jp/icl/signal/hori/publications/thori_csa2004.pdf

[2] S. Araki, H. Sawada, and M. Shoji, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. ICASSP*, vol. 1, Hawaii, 2007, pp. 41–44.

[3] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[4] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[5] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2000.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. ICASSP*, vol. 5, May 2006, pp. 33–36.

[7] R. Astudillo, D. Kolossa, and R. Orglmeister, "Propagation of statistical information through non-linear feature extractions for robust speech recognition," in *Proc. MaxEnt2007*, 2007.

[8] D. Kolossa and R. Orglmeister, "Separation and recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. WASPAA 2005*, ser. Lecture Notes in Computer Science, vol. 3195. Springer, 2005, pp. 832–839.

[9] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, May 2005.

[10] J. Barker, P. Green, and M. Cooke, "Linking auditory scene analysis and robust ASR by missing data techniques," in *Proceedings WISP 2001*, 2001.