GENERALIZATION OF SUPERVISED LEARNING FOR BINARY MASK ESTIMATION

Tobias May[*]

Technical University of Denmark
Centre for Applied Hearing Research
DK - 2800 Kgs. Lyngby, Denmark
tobmay@elektro.dtu.dk

Timo Gerkmann[†]

Speech Signal Processing Group
Cluster of Excellence Hearing4all
University of Oldenburg, Germany
timo.gerkmann@uni-oldenburg.de

ABSTRACT

This paper addresses the problem of speech segregation by estimating the ideal binary mask (IBM) from noisy speech. Two methods will be compared, one supervised learning approach that incorporates *a priori* knowledge about the feature distribution observed during training. The second method solely relies on a frame-based speech presence probability (SPP) estimation, and therefore, does not depend on the acoustic condition seen during training. We investigate the influence of mismatches between the acoustic conditions used for training and testing on the IBM estimation performance and discuss the advantages of both approaches.

Index Terms— ideal binary mask, speech segregation, generalization, speech presence probability

1. INTRODUCTION

In speech communication devices, like mobile telephones or hearing aids, the captured speech is often disturbed by additive noise. This additive noise can reduce both the speech quality and the speech intelligibility. Speech enhancement algorithms can be employed to improve speech quality, while blindly improving speech intelligibility is considered a difficult task [1, 2].

It was shown that if an ideal binary mask (IBM) or an ideal continuous gain (CG) function is available in the short-time spectral domain, intelligible speech can be obtained even at a very low signal-to-noise ratio (SNR) [3, 4]. In this paper we focus on the IBM, which is a binary matrix that classifies a time-frequency (T-F) representation of noisy speech into target-dominated and masker-dominated T-F units. To construct the IBM *a priori* knowledge about the target and the masker is required. However, while the IBM is capable of improving speech intelligibility, the IBM is not available in practice and hence has to be blindly estimated from the noisy mixture.

Considering the problem of single-channel binary mask estimation, recent studies have employed the supervised learning of amplitude modulation spectrogram (AMS) features [5, 6]. In general, classification-based approaches exploit *a priori* knowledge about the distribution of acoustic features that are extracted from speech and noise signals during an initial training stage. We will refer to these approaches as *pre-trained*. Any mismatch between the acoustic conditions that occur during training and testing can distort the observed feature distribution, which in turn is likely to degrade classification performance. Often this problem is avoided by evaluating classification-based systems only under those acoustic conditions that have been included also in the training stage [5, 6]. Therefore, one of the biggest challenges is to design classification-based segregation systems that are able to generalize to unseen acoustic conditions.

The problem of separating the contribution of speech and noise given the noisy mixture has strong similarities with speech enhancement algorithms, where typically an estimation of the background noise power spectral density (PSD) is employed to determine the SNR or the speech presence probability (SPP) in individual discrete Fourier transform (DFT) bins [7, 8, 9]. In contrast to classification-based systems, these approaches are not trained for a specific acoustic environment. Instead, rather general assumptions are made on the prior probability density function (PDF) of clean speech and the background noise. Typical examples are a Gaussian distribution for the complex noise coefficients in the short-time Fourier domain, and a Gaussian or super-Gaussian distribution for the speech coefficients [9]. We will refer to these approaches as *generic* approaches.

The aim of this study is to analyze the influence of *a priori* knowledge on speech segregation performance by means of estimating the IBM. A generic SPP-based approach for the estimation of the IBM is presented that does not rely on an initial training stage. This approach is then compared to a pre-trained supervised-learning strategy where the distribution of AMS features is learned by a Gaussian mixture model (GMM) classifier [5, 6]. During evaluation, the mismatch between the acoustic conditions used for training and testing is systematically varied and the influence on IBM estimation performance is investigated. Specifically, different forms of

[*]Supported by EU FET grant TWO!EARS, ICT-618075.
[†]Supported by German Research Foundation (DFG) Grant GE2538/2-1.

978-1-4799-6808-4/14/$31.00 ©2014 IEEE

mismatch are considered, ranging from solely spectral distortions due to band-pass filtering, to spectro-temporal smearing caused by room reverberation. In addition, the ability to segregate speech in the presence of noise that has not been seen during training is evaluated. We discuss the advantages of both methods for the task of IBM estimation.

2. SYSTEM DESCRIPTION

2.1. Ideal binary mask

The IBM requires *a priori* knowledge about the short-time DFT representation of speech $S_k(\ell)$ and noise $N_k(\ell)$, where k denotes the frequency index and ℓ indexes the time segment. If speech and noise are given, the IBM can be determined by comparing the local *a priori* SNR to a local criterion (LC)

$$\mathrm{IBM}_c(\ell) = \begin{cases} 1 & \text{if } 10\log_{10}\left(\frac{\sum_k G_{k,c}|S_k(\ell)|^2}{\sum_k G_{k,c}|N_k(\ell)|^2}\right) > \mathrm{LC} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

An element in the IBM is 1 if the local *a priori* SNR is larger than the LC and is 0 otherwise. The spectral resolution is inspired by the human auditory system, where $G_{k,c}$ reflects the frequency-dependent response of 25 auditory filters that cover a frequency range between 80 Hz and 8000 Hz according to the mel-frequency spacing [10]. We denote the index of a frequency band in the auditory domain by c. In the following, the generic SPP-based and the pre-trained classification-based approach are described in detail.

2.2. Generic speech presence probability (SPP)-based mask estimation

In the short-time DFT-domain, we observe the noisy speech $Y_k(\ell)$ as the linear superposition of speech $S_k(\ell)$ and additive noise $N_k(\ell)$. To estimate the IBM without pre-training, we first estimate the *a posteriori* SPP in the DFT-domain using [7]. The *a posteriori* SPP is defined as the probability that speech is present in a certain T-F unit, given the noisy observation $Y_k(\ell)$. Introducing the hypothesis of speech presence $\mathcal{H}_1$, this *a posteriori* SPP can be written as $P(\mathcal{H}_1 \mid Y_k(\ell))$.

The approach in [7] estimates the noise PSD based on an estimate of the *a posteriori* SPP. Given a previous estimate of the noise PSD $\widehat{\sigma}^2_{\mathrm{N},k}(\ell-1)$, under the assumption of complex Gaussian distributed speech and noise short-time DFT coefficients, the *a posteriori* SPP is obtained as

$$\begin{aligned}\mathcal{P}_k(\ell) &= P(\mathcal{H}_1 \mid Y_k(\ell)) \\ &= \left(1+(1+\xi)\exp\left(-\frac{|Y_k(\ell)|^2}{\widehat{\sigma}^2_{\mathrm{N},k}(\ell-1)}\,\frac{\xi}{1+\xi}\right)\right)^{-1}. \end{aligned} \tag{2}$$

Here, ξ represents the SNR that can be expected in speech presence. Minimizing the total probability of error, ξ corresponds to 15 dB [7]. The estimate of the noise PSD is then updated based on the *a posteriori* SPP [7]

$$\begin{aligned}|\widehat{N_k(\ell)}|^2 = (1 - P(\mathcal{H}_1 \mid Y_k(\ell)))\,|Y_k(\ell)|^2 \\ + P(\mathcal{H}_1 \mid Y_k(\ell))\,\widehat{\sigma}^2_{\mathrm{N},k}(\ell-1)\end{aligned} \tag{3}$$

and recursive smoothing

$$\widehat{\sigma}^2_{\mathrm{N},k}(\ell) = 0.8\,\widehat{\sigma}^2_{\mathrm{N},k}(\ell-1) + 0.2\,|\widehat{N_k(\ell)}|^2. \tag{4}$$

To facilitate the comparison of the generic approach and the pre-trained approach, we integrate the DFT-domain *a posteriori* SPP information $\mathcal{P}_k(\ell)$ into auditory channels by considering the frequency-dependent contribution to the overall auditory channel energy

$$\widetilde{\mathcal{P}}_c(\ell) = \frac{\sum_k G_{k,c}\mathcal{P}_k(\ell)\,|Y_k(\ell)|^2}{\sum_k G_{k,c}|Y_k(\ell)|^2}. \tag{5}$$

In the context of SPP estimation [11] as well as IBM estimation [6] it has been shown that the exploration of information present in neighboring T-F units is beneficial. Therefore, we follow the approach presented in [6] and integrate the estimated *a posteriori* SPP in the auditory domain $\widetilde{\mathcal{P}}_c(\ell)$ across a plus-shaped neighborhood function that spans over 5 auditory channels and 5 time frames.

Finally, an estimate of the IBM is obtained by applying a noise-specific threshold Ψ to the estimated SPP, as

$$\widehat{\mathrm{IBM}}_c(\ell) = \begin{cases} 1 & \text{if } \widetilde{\mathcal{P}}_c(\ell) > \Psi \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

2.3. Classification-based mask estimation

The classification-based approach exploits *a priori* knowledge about the feature distribution of speech-dominated and noise-dominated T-F units during an initial training stage.

Firstly, the AMS features are extracted according to [6]. Therefore, the noisy speech is normalized according to its root mean square (RMS) value and then divided into overlapping frames of 4 ms duration with a shift of 0.25 ms. Each frame is Hamming windowed and zero-padded to a length of 128 samples and a 128-point DFT is computed. The DFT magnitudes are multiplied by 25 auditory filters that covered a frequency range between 80 Hz and 8000 Hz according to the mel-frequency spacing [10]. The envelope in each auditory filter is extracted by full-wave rectification and further analyzed for segments of 32 ms duration by 15 triangular-shaped modulation filters that are linearly-spaced between 15.6 and 400 Hz, resulting in a set $\boldsymbol{X}_c(\ell)$ of 15 AMS features $M_c(\ell)$ for each auditory channel and each time frame, as $\boldsymbol{X}_c(\ell) = \{M_c^1(\ell), \ldots, M_c^{15}(\ell)\}^T$. In this contribution we extend the AMS feature extraction, as described in [5, 6], by a normalization stage. More specifically, we propose to normalize the envelope in each auditory channel by its median prior to modulation analysis. The required normalization statistics were

computed over the entire sentence. This new normalization stage aims at reducing the influence of the overall background noise level on the AMS feature distribution, which is expected to improve the generalization of the pre-trained approach to unseen acoustic conditions.

Secondly, a two-class Bayesian classifier is trained to distinguish between speech-dominated and noise-dominated T-F units for each auditory channel, denoted as λ_1^c and λ_0^c respectively. Given the AMS feature vector $\boldsymbol{X}_c(\ell)$, the IBM is estimated by comparing two *a posteriori* probabilities [5, 6]

$$\widehat{\mathrm{IBM}}_c(\ell) = \begin{cases} 1 & \text{if } P(\lambda_1^c|\boldsymbol{X}_c(\ell)) > P(\lambda_0^c|\boldsymbol{X}_c(\ell)) \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Following [6], the IBM estimation can be improved by exploiting the *a posteriori* probability of speech and noise presence, $P(\lambda_1^c|\boldsymbol{X}_c(\ell))$ and $P(\lambda_0^c|\boldsymbol{X}_c(\ell))$, over neighboring T-F units. Similar to the SPP-based approach, we use a plus-shaped neighborhood function that spans over 5 auditory channels and 5 adjacent time frames.

3. EVALUATION

Signals are processed in 32 ms frames with 50 % overlap at a sampling frequency of $f_s = 16$ kHz. Noisy speech with an average length of 3 s is created by corrupting randomly selected male and female sentences from the TIMIT database [12] with a randomly chosen excerpt of one of four background noises listed in Tab. 1. The speech and noise corpora are split in two halfs of equal size to ensure that there is no overlap between the signals used for training and testing.

The GMM-based classification system is trained with 60 sentences at -5, 0 and 5 dB SNR. During training, a LC of -5 dB is used to separate the AMS features into speech-dominated and noise-dominated elements based on the local *a priori* SNR. For each background noise and auditory channel, a separate GMM classifier is trained with 16 Gaussian components and full covariance matrices. To analyze the benefit of the feature normalization stage and the exploration of neighboring T-F units, we evaluate three GMM classifiers: *GMM1* denotes a GMM classifier based on AMS features, *GMM2* is a GMM classifier based on normalized AMS features and *GMM3* refers to a GMM classifier based on normalized AMS features which exploits the plus-shaped neighborhood function as proposed in [6]. Furthermore, *SPP1* denotes the SPP-based IBM estimation and *SPP2* additionally exploits the plus-shaped neighborhood function.

The speech segregation performance is measured by comparing the estimated binary mask with the IBM. Specifically, the percentage of correctly classified speech-dominated T-F units (HIT rate) and the percentage of erroneously identified noise-dominated T-F units (FA rate) are computed. As proposed in [5] we report their difference, HIT - FA, as a measure of the overall segregation performance. The evaluation is based on 60 sentences mixed at -5 and 0 dB SNR. To study the influence of mismatches between the acoustic conditions used for training and testing, the pre-trained and the generic systems are evaluated under four conditions:

1. Matched: no mismatch between training and testing.
2. BP: speech and noise signals are band-limited to the frequency range $1000-3500$ Hz by a second-order butterworth band-pass filter to simulate effects similar to that of a narrow band telephone transmission channel.
3. BRIR: speech and noise signals are convolved with a binaural room impulse response (BRIR) corresponding to *Room A* [15], with a reverberation time of $T_{60} = 0.32$ s and a direct-to-reverberant ratio (DRR) of 6.06 dB. The azimuth is randomly selected for each mixture and the impulse response corresponding to the better ear is chosen.
4. Noise: for a given background noise, the segregation systems are trained with the three background noises that are not used for evaluation.

Table 1. Types of background noises.

Noise type	Description	Sec.
ICRA1 [13]	Stationary speech-shaped noise	120
ICRA7 [13]	Speech-shaped and modulated noise	1200
Factory [14]	Noise inside a car production hall	235
Cockpit [14]	F-16 traveling at 500 knots	235

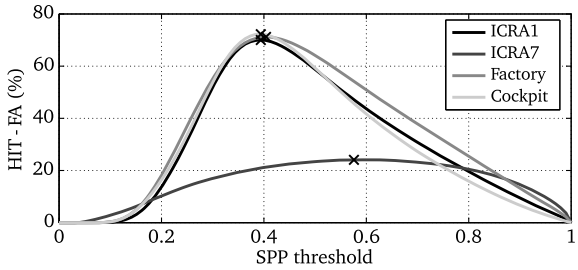

Fig. 1. Noise-specific selection of the SPP threshold Ψ used for the IBM estimation in (6).

The only training aspect of the generic approach is the selection of the noise-specific threshold Ψ, employed in (6), that maximizes the HIT - FA on the training data, as illustrated in Fig. 1 for *SPP2*. Best results are obtained when a conservative threshold is used. For ICRA7, a flat curve without a distinct maximum is observed, which indicates that the SPP-based approach is not able to track the non-stationary multi-talker noise. The shapes for the other three noise types ICRA1, factory and cockpit are very similar to each other, suggesting that an optimal threshold might be generally applicable.

4. EXPERIMENTAL RESULTS

The speech segregation performance is shown in Fig. 2 as a function of the mismatch between the training and test-

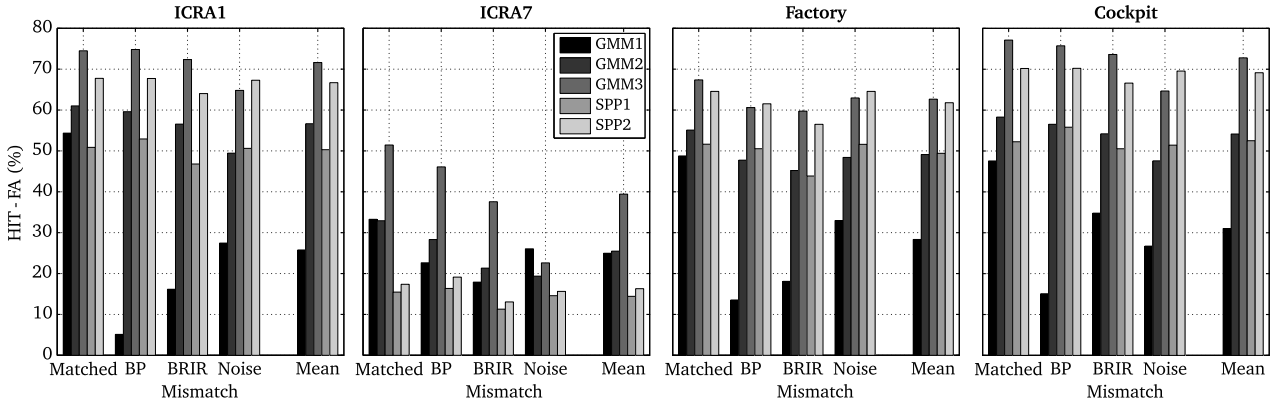

Fig. 2. Segregation performance of individual T-F units as a function of the mismatch between training and testing conditions.

ing conditions. The pre-trained system *GMM1* only performs well in the matched condition. As soon as spectral distortions or spectro-temporal smearing occur in the testing stage (e.g. due to band-pass filtering or reverberation), the distribution of the extracted AMS features does not match with the distribution learned during training, which prevents the model from performing well under mismatched conditions. The normalization stage, as described in Sec. 2.3, additionally included in *GMM2* can compensate for this mismatch, and consequently, the performance under mismatched conditions is almost as good as in the matched condition.

The performance of the SPP-based approach *SPP1* is not very sensitive to any of the mismatch conditions. For the speech-modulated ICRA7 noise, the classification-based IBM estimation is superior to the SPP-based approach, most noticeably when the acoustic conditions between training and testing match. However, this advantage reduces with increasing mismatch between training and testing conditions. Considering ICRA1, factory and cockpit noise, there is a slight advantage for the SPP-based approach when the corresponding background noise has not been seen during training (mismatch condition *Noise*). In addition, the observation that the optimal SPP threshold Ψ is very similar for three different noises (except for ICRA7) facilitates the application of the SPP-based approach to a wide range of background noises.

In general, the exploration of context information, as performed by *GMM3* and *SPP2*, is beneficial for both the *pre-trained* and the *generic* approaches. However, this is only true if the extracted properties of the speech and noise signal match with the model expectation, which is not the case for the ICRA7 noise, in particular in the mismatched noise condition.

5. DISCUSSION AND CONCLUSIONS

We have considered the problem of estimating the IBM from noisy speech mixtures. For this purpose, the supervised learning of AMS features (pre-trained) was compared to a frame-based estimation of speech presence probability (generic). It was shown that pre-trained approaches can be quite sensitive with respect to a mismatch between training and testing conditions, which may occur when the noisy speech is processed by a telephone transmission channel or modified by room reverberation. In contrast, the generic approach was robust against these mismatches, and the achieved segregation performance was similar to the pre-trained approach, despite the fact that no *a priori* information was required. In addition, the generic approach has the ability to generalize to unseen acoustic conditions, including unseen background noises. However, for highly non-stationary noises, like the speech-modulated ICRA7 noise, the performance of the generic approach is limited and outperformed by the pre-trained approach with respect to HIT - FA.

To improve the robustness of pre-trained approaches to mismatches between training and testing conditions, a median-based normalization technique was proposed. The normalized AMS features greatly improved segregation performance under both matched and mismatched conditions. Whereas spectral coloration can be dealt with quite effectively, the presence of unseen noises will alter the observed AMS feature distribution, which reduces the performance of pre-trained approaches. This performance degradation was most noticeable for the speech-modulated ICRA7 noise, which exhibited a very specific AMS feature distribution.

The consideration of contextual information about speech activity was beneficial for both the generic and the pre-trained approaches and substantially improved the accuracy of the IBM estimation.

Finally, the feature space of classification-based approaches can be readily extended with additional features that can further improve the distinction between speech and noise activity. However, the effectiveness of any additional feature should be evaluated under mismatched conditions.

6. REFERENCES

[1] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1491–505, 2010.

[2] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[3] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[4] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, 2012.

[5] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[6] T. May and T. Dau, "Environment-aware ideal binary mask estimation using monaural cues," in *Proc. WASPAA*, 2013, pp. 1–4.

[7] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[8] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 108–121, 2012.

[9] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-art*, Morgan & Claypool, Colorado, USA, Feb. 2013.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[11] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)*, 1993.

[13] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

[14] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[15] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, 2010.