INTEGRATION OF BEAMFORMING AND AUTOMATIC SPEECH RECOGNITION THROUGH PROPAGATION OF THE WIENER POSTERIOR

Ramón Fernandez Astudillo, Alberto Abad and João Paulo da Silva Neto

Spoken Language Laboratory, INESC-ID-Lisboa, Lisboa, Portugal

ramon@astudillo.com,{alberto.abad,Joao.Neto}@inesc-id.pt

ABSTRACT

This paper details one of the front-end components of the system used at the PASCAL-CHiME multi-source robust automatic speech recognition (ASR) challenge 2011. The presented approach uses uncertainty propagation techniques to integrate conventional beamforming with automatic speech recognition. The paper addresses the derivation of a complex Gaussian posterior for the multi-channel Wiener and the delay and sum beamformer and introduces a new approach based on the propagation of the Wiener posterior through the resynthesizing process. Results on the PASCAL-CHiME task for this algorithms show that they consistently outperform conventional beamfomers with a minimal increase in computational complexity.

Index Terms— Beamforming, Uncertainty Propagation, Observation Uncertainty, Robust ASR

1. INTRODUCTION

In the past decade a new approach to robust ASR has emerged that is aimed to propagate the uncertainty in the acoustic features due to either the noise effect itself or the residual noise [1]. Such an approach allows the dynamic compensation of the acoustic models offering interesting trade-offs in terms of robustness and computational complexity. Short-time Fourier (STFT) uncertainty propagation [2] attempts to present a generic framework for the integration of conventional speech enhancement in STFT domain and ASR through observation uncertainty techniques. Such an integration is attained by considering the STFT after speech enhancement as a random variable rather than a deterministic point estimate. This uncertain description of the spectrum is then transformed, propagated, into the feature domain yielding a posterior distribution of the features. This posterior can be combined with observation uncertainty techniques like front-end uncertainty decoding [3] or modified imputation [4] for a more robust ASR.

One of the advantages of uncertainty propagation is that it can be very easily extended to incorporate expertise of the speech enhancement field. This paper explores in particular the integration of two well established multichannel speech enhancement techniques, the delay and sum beamformer [5] and the multichannel Wiener filter [6] with observation uncertainty techniques. The detailed algorithms form part of the robust multi-source ASR system presented to the PASCAL-CHiME challenge 2011 [7] and are here extended to support an intermediate resynthesizing step, which further broadens the range of applicable speech enhancement techniques.

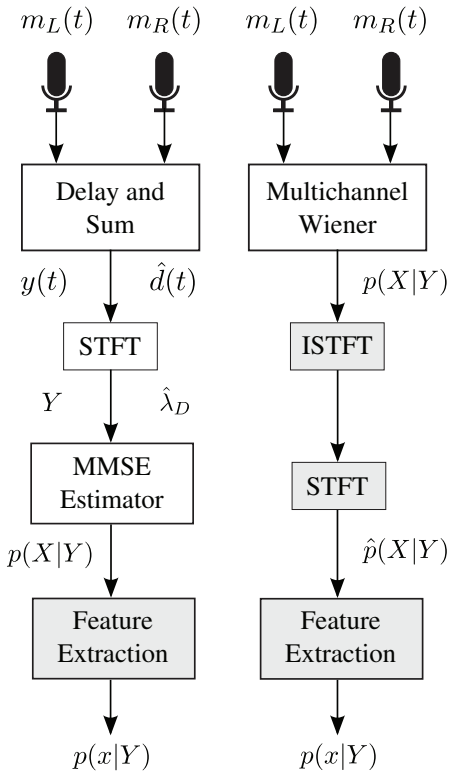

Fig. 1. Left: A noise variance estimate is produced at the beamforming stage, the STFT posterior is generated with a single channel MMSE estimator. Right: The STFT posterior is generated directly at the beamforming stage and propagated through the ISTFT and STFT. Shaded: Stages through which there is propagation.

This work is supported by the Portuguese Foundation for Science and Technology, grant number SFRH/BPD/68428/2010

The paper is divided as follows Sections 2 and 3 discuss the estimation of uncertainty for delay and sum and multichannel Wiener beamformers respectively. Section 4 reviews briefly the propagation of a posterior distribution of the spectrum for the selected feature extraction and recognition under observation uncertainty. Section 5 introduces tests on the PASCAL-CHiME task and analyzes the results. Finally Section 6 presents the conclusions.

2. UNCERTAINTY IN A DELAY AND SUM BEAMFORMER

The delay-and-sum (DS) beamformer aligns the different microphone signals to compensate for the different path lengths from the source to the various microphones. For the proposed scenario it is assumed that the speaker lies directly in front of a pair of microphones and thus the DS beamformer can be reduced to a simple addition of the channels

$$y(t) = m_L(t) + m_R(t), \tag{1}$$

where $m_L(t)$ and $m_R(t)$ correspond to the left and right microphone signals. Since such a simple spatial filter can only partially suppress directional noises, it is often complemented with a second speech enhancement stage. A simple approach is to employ minimum mean square error (MMSE) estimators with a complex Gaussian model like the Wiener or Ephraim-Malah filters [8]. Let Y be a single Fourier coefficient of the STFT of $y(t)$, this model assumes that Y is the sum of two hidden random variables X and D, corresponding to the Fourier coefficients of speech and noise respectively. These are assumed to be circularly symmetric complex Gaussian random variables with zero mean and variances λ_X and λ_D respectively. Applying the Bayes theorem for this model, the following complex Gaussian posterior distribution of the clean STFT is obtained

$$p(X|Y) = \mathcal{N}_{\mathbb{C}}\left(\hat{X}^W, \lambda\right), \tag{2}$$

where the mean

$$\hat{X}^W = G \cdot Y = \frac{\lambda_X}{\lambda_X + \lambda_D} Y \tag{3}$$

is the Wiener estimator of the clean Fourier coefficient and the variance

$$\lambda = \frac{\lambda_X \lambda_D}{\lambda_X + \lambda_D} \tag{4}$$

is the residual mean square error (MSE) [9]. This posterior yields, in fact, an uncertain description of the clean spectrum given the available information and thus can be used for uncertainty propagation. The only question remaining is how to determine the a priori variances of each speech and noise Fourier coefficients λ_X and λ_D. Setting these values using the available information from the beamforming step effectively allows to integrate this pre-processing step with the rest of the front-end. The limitations of the DS beamformer as a frequency-space filter are well known, however, in a environment with possible directional noises and reverberance it is very difficult to find a formulation for the estimation uncertainty. Since the speaker is situated in front of the two microphones, a simple approach is to consider any asymmetry between the microphone signals as uncertainty about the observed spectrum Y[1]. One simple measure of asymmetry can be obtained as

$$\lambda_D = \left|\text{STFT}\left\{\hat{d}(t)\right\}\right|^2 = \left|\text{STFT}\left\{m_L(t) - m_R(t)\right\}\right|^2. \tag{5}$$

The speech variance λ_X can be then determined by a suitable method like the decision directed method [8, Eq. 51].

3. UNCERTAINTY IN A MULTICHANNEL WIENER FILTER

Rather than using single channel MMSE estimators as a post-processing stage to a DS beamfomer, a multi-channel Wiener filter can be used. The most common assumption used to derive this filter is that the noise present in both microphones has low correlation compared to speech. This leads to the following approximation of the Wiener gain

$$G = \frac{\lambda_X}{\lambda_X + \lambda_D} \approx \frac{2 \cdot E\{M_L M_R^*\}}{E\{|M_L|^2\} + E\{|M_R|^2\}} \tag{6}$$

where M_L and M_R correspond to the Fourier coefficients of the left and right microphone signals and $*$ to the complex conjugate operator. In practice the expectations are approximated by smoothing averages and special corrections are needed for the complex values arising from the cross-moment. For this particular implementation the real component was selected and negative values were set to zero[2].
In order to determine the parameters of the posterior in Eq. 2, the value of the noise variance λ_D needs to be determined as well. Using the same assumption as for the Wiener gain, λ_D can be estimated using the power subtraction estimator

$$\lambda_D = \frac{1}{2}\left(E\{|M_L|^2\} + E\{|M_R|^2\}\right)(1 - G) \tag{7}$$

In principle, once the parameters of the posterior have been determined it could be passed directly to the feature extraction as in [2]. However, as displayed in Fig. 1 right the multichannel Wiener filter is usually followed by a resynthesizing and posterior STFT. This operation has a notable enhancement effect since it smooths out the artifacts caused by

[1] Note that for a zero mean noise prior λ_D is related to how uncertain is that the observed coefficient Y corresponds to X.

[2] This configuration gave the best results during the PASCAL-CHiME challenge preparation.

the imperfect estimation in Eq. 6. In order to compute the posterior after ISFTF and STFT, the solution proposed for the PASCAL-CHiME Challenge was to derive the Wiener filter parameters from the Equivalent gain after these transformations [7]. Although effective, this is an ad hoc solution. This paper introduces an alternative approach in which the posterior obtained from the Wiener filter is propagated through the ISTFT and STFT transforms as in Fig. 1, right. This can be attained by joining all the linear operations involved in the ISTFT and STFT, including windowing and complex conjugate operations, into one single matrix[3] If the correlations induced by the intermediate operations like overlap and add are ignored, a simple linear closed form solution for the propagated variances can be attained.

4. UNCERTAINTY PROPAGATION AND DECODING

For a complex Gaussian posterior as that of Eq. 2 there exist very fast closed form solutions for the propagation through the Mel-Cepstral, perceptual linear prediction and other feature extractions [2]. In all these cases the resulting posterior at feature domain is well approximated by a Gaussian distribution

$$p(x|Y) \approx \mathcal{N}\left(\mu_x, \Sigma_x\right) \tag{8}$$

In case of propagating the posterior associated to a Wiener filter it can be also demonstrated that μ_x is an approximate MMSE estimator of the features and Σ_x the corresponding estimation variance [10]. This variance can be then utilized with observation uncertainty techniques to further improve ASR robustness. The most often used approach for this matter is front-end uncertainty decoding (UD), which simply adds the uncertainty variances Σ_x to the acoustic model variance [3]. An alternative approach, which consistently shows a better performance in combination with uncertainty propagation, is modified imputation (MI) [4] which re-estimates the features based on the uncertainty to acoustic model variance ratio as

$$\hat{x} = \frac{\Sigma_q}{\Sigma_q + \Sigma_x}\mu_x + \frac{\Sigma_x}{\Sigma_q + \Sigma_x}\mu_q, \tag{9}$$

where μ_q and Σ_q correspond to the mean and variance of each mixture. MI also has lower computational complexity than UD since it does not require the re-computation of the mixture determinant.

5. EXPERIMENTAL RESULTS

In order to test the efficiency of the proposed approaches the PASCAL-CHiME challenge task was used. This recently created corpus [11] provides a realistic scenario for binaural automatic speech recognition in room environments. The task is derived from the GRID corpus by convolving clean speech with real room impulse responses as well as adding various noise sources at different spatial localizations. Train data is clean but reverberated and the final test set includes a slightly different room impulse response. The feature extraction used was amplitude based Mel-Cepstral coefficients with cepstral mean subtraction. As recognition engine a modified version of HTK capable of performing uncertainty decoding and modified imputation was used. The HTK training, testing and scoring scripts provided in the challenge were used.

For the case of the DS beamfomer using channel asymmetry as noise estimate, conventional WIENER and MMSE estimators of the amplitude (MMSE-STSA) and log amplitude (MMSE-LSA) were compared against uncertainty propagation (MMSE-MFCC) with and without MI and UD. The multichannel Wiener approach was compared with uncertainty propagation also with MI and UD alternatives. Two forms of determining uncertainty where compared, the equivalent gain after STFT proposed in [7] (EQWIN) and the propagation through ISTFT proposed in this paper (IDFTUP).

Table I displays the results of the DS experiments. As expected the use of MMSE post-processing greatly reduces the Word Error Rate (WER) of the baseline with MMSE-LSA showing the best performance on average. The aggressiveness of the Wiener filter also provides a positive trade-off for low SNRs. When used alone, the MMSE-MFCC estimator provides a performance lower than that of the MMSE-LSA but when combined with UD and particularly MI it outperforms all other methods. The results of the multichannel Wiener experiments are displayed in Table II and present a similar tendency. The use of UD and MI considerably improve the baseline values although the multichannel Wiener is far more effective than the DS in removing distortions. The proposed IDFTUP fails to outperform the previous uncertainty estimation method although it presents a better behavior at low SNR. A reason for this could be that proposed solution is an approximate propagation of the MSE, which does not include errors in the estimation of a priori parameters λ_D or λ_X. The ad hoc solution on [7] might however partially compensate for this fact.

6. CONCLUSIONS

Two possible approaches for the integration of beamforming and observation uncertainty techniques have been analyzed in detail and a new approach that allows an intermediate resynthesizing step has been introduced. The proposed methods show a notable improvement over conventional beamforming approaches with a low increase in computational complexity. The estimation of uncertainties at the beamforming stage need however to be further studied to include errors in the estimation of a priori information.

[3] In practice this is separated into various matrices for real and imaginary components. Pre-emphasis was also ignored since it excessively increased high frequency variances.

Table 1. WER [%] and relative error reduction with respect to baseline for the delay and sum beamformer

	-6db	-3db	0dB	3dB	6dB	9dB	∞	r.r.[%]
Delay and Sum+MMSE-MFCC+MI	57.2	**47.1**	36.2	**23.4**	**15.5**	**9.9**	5.8	**-28.4**
Delay and Sum+MMSE-MFCC+UD	58.3	48.9	37.2	24.8	16.7	10.3	6.3	-24.8
Delay and Sum+MMSE-LSA	59.2	50.9	38.7	26.2	17.6	10.7	5.8	-23.7
Delay and Sum+MMSE-MFCC	59.6	50.1	38.2	25.7	18.2	10.8	6.4	-22.2
Delay and Sum+MMSE-STSA	59.4	53.0	42.6	27.9	18.7	11.8	**5.2**	-21.1
Delay and Sum+Wiener	**53.9**	**47.1**	**34.6**	24.5	18.4	13.4	7.6	-20.6
Delay and Sum	66.3	58.7	45.9	34.0	22.1	13.8	5.6	-10.6
Baseline (single channel)	68.4	62.6	51.7	37.6	25.8	16.2	6.6	0

Table 2. WER [%] and relative error reduction with respect to baseline for the multichannel Wiener

	-6db	-3db	0dB	3dB	6dB	9dB	∞	r.r.[%]
Multi-Channel Wiener+EQWIN+MMSE-MFCC+MI	52.9	44.2	31.4	**20.2**	13.8	**9.3**	**5.5**	**-34.7**
Multi-Channel Wiener+IDFTUP+MMSE-MFCC+MI	**51.7**	**42.5**	**30.0**	**20.2**	**13.4**	9.7	6.8	-32.9
Multi-Channel Wiener+IDFTUP+MMSE-MFCC+UD	53.8	43.1	30.4	21.8	13.8	10.6	7.0	-30.1
Multi-Channel Wiener+EQWIN+MMSE-MFCC+UD	54.5	45.8	33.3	22.2	14.6	10.5	6.4	-29.2
Multi-Channel Wiener+IDFTUP+MMSE-MFCC	53.7	44.2	30.9	22.3	14.2	10.6	7.1	-29.1
Multi-Channel Wiener+EQWIN+MMSE-MFCC	56.1	48.2	34.5	23.6	15.8	10.8	6.9	-25.5
Multi-Channel Wiener	56.6	48.5	33.9	22.1	15.3	11.4	7.2	-25.0
Baseline (single channel)	68.4	62.6	51.7	37.6	25.8	16.2	6.6	0

7. REFERENCES

[1] L. Deng, "Front-end, back-end, and hybrid techniques to noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds., chapter 4, pp. 67–99. Springer, Berlin, Germany, 2011.

[2] Ramon Fernandez Astudillo, *Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition*, Ph.D. thesis, Technical University Berlin, 2010.

[3] N.B. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Trans. Speech, Audio Processing*, vol. 10 (3), pp. 158–166, 2002.

[4] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.

[5] Don H. Johnson and Dan E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, 1993.

[6] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Acoustics, Speech, and Signal Processing, International Conference on*, apr 1988, vol. 5, pp. 2578 –2581.

[7] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. Neto, and R. Martin, "Chime challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *Int. Workshop on Machine Listening in Multisource Environments*, 2011.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[9] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Proc. Interspeech*, 2009, pp. 2491–2494.

[10] R. F. Astudillo and R. Orglmeister, "A MMSE estimator in mel-cepstral domain for robust large vocabulary automatic speech recognition using uncertainty propagation," in *Proc. Interspeech*, 2010, pp. 713–716.

[11] H. Christensen, J. Barker, N. Ma, and P. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010, pp. 1918–1921.