A MMSE Estimator in Mel-Cepstral Domain for Robust Large Vocabulary Automatic Speech Recognition using Uncertainty Propagation

Ramón Fernandez Astudillo, Reinhold Orglmeister

Chair of Electronics and Medical Signal Processing, Technical University Berlin, Germany

ramon@astudillo.com, Reinhold.Orglmeister@tu-berlin.de

Abstract

Uncertainty propagation techniques achieve a more robust automatic speech recognition (ASR) with modeling the information missing after speech enhancement in the short-time Fourier transform (STFT) domain in probabilistic form. This information is then propagated into the feature domain where recognition takes place and combined with observation uncertainty techniques like uncertainty decoding. In this paper we show how uncertainty propagation can also be used to yield minimum mean square error (MMSE) estimates of the clean speech directly in the recognition domain. We develop a MMSE estimator for the Mel-cepstral features by propagation of the Wiener filter posterior distribution and show how it outperforms conventional MMSE methods in the STFT domain on the AU-ROA4 large vocabulary test environment.

Index Terms: Wiener Filter, Ephraim-Malah filters, MMSE, Uncertainty Propagation, Modified Imputation, Large Vocabulary ASR

1. Introduction

The robustness of automatic speech recognition (ASR) systems can be increased by employing model based speech enhancement techniques like [1], which dynamically adapt the ASR models to noisy environments. Such methods require however of enough available noise samples and have usually high computational costs. An inexpensive alternative to dynamic model adaptation is to employ conventional speech enhancement techniques, which are used to improve the quality of corrupted speech signals for human listeners. Conventional speech enhancement techniques, like the Wiener or Ephraim-Malah filters, operate in the short-time Fourier transform (STFT) domain where the interaction between speech and noise is easier to model. Automatic speech recognition (ASR) systems work however in feature domains non-linearly related to the STFT domain which facilitates the learning and classification of speech. The direct use of STFT domain speech enhancement methods for robust ASR is therefore sub-optimal. To achieve a better integration of STFT domain speech enhancement with ASR systems, uncertainty propagation techniques might be used [2, 3, 4].

Uncertainty propagation techniques consider the signal after speech enhancement as ‘uncertain’ or probabilistic rather than deterministic, where this uncertainty represents the information missing after speech enhancement. This probabilistic description of the clean signal is then non-linearly transformed into the feature domain where the ASR system is trained. The resulting probabilistic description of the speech features can be used to obtain a more robust speech recognition by means of observation uncertainty techniques like modified imputation [2] or uncertainty decoding [5].

The complex Gaussian uncertainty model [4] provides a simple representation of each uncertain Fourier coefficient and results in an approximately Gaussian distribution for some of the most used feature extractions like the Mel-Cepstral features or the RASTA-LPCs. Such models have been successfully used to improve the performance of both multi-channel and single-channel speech enhancement algorithms like ICA with post-masking [6] or the advanced front-end [7]. Particularly interesting is the case in which speech enhancement is attained with conventional minimum-mean square error (MMSE) estimators in the STFT domain. In this case, exact and approximate closed form solutions can be developed to determine the resulting uncertainty [8]. Furthermore, as it will be shown in this paper, in the particular case of the propagation of the Wiener estimator, uncertainty propagation can be used to provide a MMSE estimate in feature domain. Since this estimate is directly computed in the domain of recognition features it yields superior performance to that of the conventional MMSE techniques in Fourier, amplitude or log-amplitude domains. Furthermore the variance of the estimator is also available, thus providing a measure of uncertainty which can be used to improve recognition robustness by means of observation uncertainty techniques. The efficiency of the proposed estimator and its combination with observation uncertainty techniques is tested for large vocabulary speech recognition.

2. MMSE Speech Enhancement

Let \( y(t) \) be a noisy version of a clean speech signal \( x(t) \) and let \( Y_{kl} \) denote a time-frequency representation of \( y(t) \) obtained by applying the STFT. Each Fourier coefficient of the STFT of \( y(t) \) can be assumed to contain a mixture of the Fourier coefficient of the original clean signal \( X_{kl} \) and the Fourier coefficient of noise \( D_{kl} \) given as

\[
Y_{kl} = X_{kl} + D_{kl}. \tag{1}
\]

MMSE speech enhancement methods in the STFT domain aim to obtain an estimation of the spectrum of the clean signal by minimizing the expected estimation error. This requires of establishing an a priori distribution of the hidden Fourier coefficients \( X_{kl} \) and a likelihood function linking observable and hidden Fourier coefficients \( Y_{kl} \) and \( X_{kl} \). The most extended statistical model for this purpose is to assume that all Fourier coefficients are statistically independent and that speech and noise are zero mean complex Gaussian distributed. This leads to the individual modeling of each unknown clean Fourier coefficient with following prior distribution

\[
p(X) = \frac{1}{\pi \lambda X} \exp \left( -\frac{|X|^2}{\lambda X} \right) \tag{2}
\]
where the frequency bin \( k \) and frame \( l \) indices have been removed for simplicity. The parameter \( \lambda_X \) corresponds to the variance of the Fourier coefficient \( X \). Similarly we can define the likelihood distribution for each pair of corresponding noisy and clean Fourier coefficients \( Y \) and \( X \) as

\[
p(Y|X) = \frac{1}{\pi \lambda_D} \exp \left( -\frac{|Y - X|^2}{\lambda_D} \right)
\]

where \( \lambda_D \) corresponds to the variance of noise. Once the model has been defined, an estimation \( \hat{X} \) of each clean Fourier coefficient \( X \) which minimizes the mean square error (MSE) can be obtained by solving

\[
\hat{X}_{\text{MMSE}} = \arg \min_{\hat{X}} \left\{ E \left\{ ||X - \hat{X}||^2 \right\} \right\}
\]

where \( || \) is the euclidean norm, and the expectation is computed with respect to the joint distribution of hidden \( X \) and observable data \( Y \). It is easy to see that the solution to a MMSE problem corresponds always to the center of probabilistic density or mean of the posterior distribution, in this case

\[
\hat{X}_{\text{MMSE}} = E\{X|Y\}
\]

In the particular case of the Gaussian model, not only the mean but the whole posterior distribution is easy to obtain and results in the complex Gaussian distribution

\[
p(X|Y) = \frac{1}{\pi \lambda} \exp \left( -\frac{|X - \hat{X}^W|^2}{\lambda} \right)
\]

where

\[
\hat{X}^W = E\{X|Y\} = \frac{\lambda_X}{\lambda_X + \lambda_D} Y
\]

is the Wiener estimator for that Fourier coefficient and

\[
\lambda = \frac{\lambda_X \lambda_D}{\lambda_X + \lambda_D}
\]

is the residual MSE of the estimator [8]. Estimating directly the clean Fourier coefficient is however sub-optimal. Estimators based on non-linear transformations of \( X \), such as the amplitude or log-amplitude usually provide superior speech enhancement performance both for ASR systems and human listeners. This is due to the fact that amplitude and log-amplitude are perceptually more relevant and also better related to the feature domains used for speech recognition. These and other similar estimators can be represented by defining a generic non-linear transformation of a Fourier coefficient

\[
\theta = f(X)
\]

and the corresponding MMSE problem

\[
\hat{\theta}_{\text{MMSE}} = \arg \min_{\hat{\theta}} \left\{ E \left\{ ||\theta - \hat{\theta}||^2 \right\} \right\} = E\{\theta|Y\} = E\{f(X)|Y\}
\]

where the fact that \( \theta \) and \( Y \) belong to different feature domains does not change the outcome of the solution. Examples of this type of MMSE estimators are the Ephraim-Malah filters in which \( \theta \) corresponds to the short-time spectral amplitude (STSA) [9] and log-spectral amplitude (LSA) [10].


diagram

Figure 1: Uncertainty propagation. After speech enhancement the speech features are considered complex Gaussian distributed. After propagation this yields Gaussian distributed features in cepstral domain.

other estimators fall however into this category, like for example when \( \theta \) corresponds to the squared short-time spectral amplitude [11] and other non-linear transformations. Interestingly, this generic type of estimators are characterized by the posterior distribution \( p(f(X)|Y) \), which can be seen as the result of propagating the complex Gaussian posterior of the Wiener filter in Eq 6 through the non-linear transformation \( f() \). As displayed in Fig. 1 this is a particular case of the problem solved by uncertainty propagation algorithms in [4]. By setting \( f() \) equal to the feature extraction transformation, we can employ uncertainty propagation to compute the expectation in Eq. 10 and achieve a MMSE estimator directly in the feature domain where recognition takes place. Rather than using uncertainty propagation only during recognition as in [8, 7], the propagated mean is then incorporated as part of the speech enhancement system and used during training.

3. MMSE Estimation in the Feature Domain of ASR

Since the feature extractions used for ASR include joint transformations between features of the same frame or between features of different time frames, the non-linear transformation has to be redefined for vector based inputs\(^1\)

\[
\theta = f(X)\]

where now \( \theta \) is the speech feature matrix and \( X \) denotes the whole STFT of the signal and not an individual Fourier coefficient. The MMSE estimator of \( \theta \) can then be obtained similarly to Eq. 10 as

\[
\hat{\theta}_{\text{MMSE}} = \arg \min_{\hat{\theta}} \left\{ E \left\{ ||\theta - \hat{\theta}||^2 \right\} \right\} = E\{f(X)|Y\}
\]

\(^1\)Feature matrices are here reshaped to vectors to simplify the notation
where $Y$ now denotes the whole STFT of the noisy signal and the posterior distribution $p(Y|X)$ has to be computed. The problem of uncertainty propagation, that is the problem of non-linearly transforming a probability distribution, can be solved in many ways. However, the strict computational cost and speed constraints of ASR reduces the number of usable techniques to a few. In [3] regression trees are trained offline to compute the transformed uncertainties whereas in [2] a mixture of closed form solutions and pseudo-Montecarlo methods are used. For the particular case of the complex Gaussian uncertainty model, solutions for the propagation through the Mel-cepstra, RASTA-LPCCs and other transformations involved in the ETSI advanced front-end are given in [4]. From this set of solutions, the conventional Mel-cepstral coefficients (MFCCs) were selected as feature extraction since they are very well known and lead to a compact closed-form solution for propagation. The Mel-cepstral coefficients are computed as the non-linear transformation

$$\theta_{il} = f(X_l) = \sum_{j=1}^{J} T_{ij} \log \left( \sum_{k=1}^{K} W_{jk} |X_{lk}|^2 \right)$$

(13)

where $X_l$ is the $l^{th}$ STFT frame and $W_{jk}$ the Mel-filterbank and discrete cosine transform (DCT) weights respectively.

For a complex Gaussian uncertainty model in STFT domain the resulting uncertainty in cepstral domain approximates a Gaussian distribution as displayed in Fig. 2

$$p(\theta_{il}|Y_l) \approx N(\mu_{\text{CEPS}}^{il}, \Sigma_{\text{CEPS}}^{il})$$

(14)

If the non-diagonal covariance after the filterbank processing is ignored, the propagated mean remains very accurate whereas the variance is underestimated. This has however no negative or even a positive effect when observation uncertainty techniques are used and it reduces the computational costs [7]. The moments of the posterior distribution in Eq. 14 can be computed by matching the second and fourth order moments of a Rice distribution with those of a log-normal distribution [7]. This results in a Gaussian distributed posterior in log-Mel domain with variance

$$\Sigma_{\text{LLG}} \approx \log \left( \frac{\sum_{k=1}^{K} W_{jk}^2 \left( 2 \lambda_{kl} |\hat{X}_{lk}|^2 + \lambda_{kl}^2 \right)}{\left( \sum_{k=1}^{K} W_{jk} \left( \lambda_{kl} + |\hat{X}_{lk}|^2 \right) \right)^2} + 1 \right)$$

(15)

and mean

$$\mu_{\text{LLG}} \approx \log \left( \sum_{k=1}^{K} W_{jk} \left( \lambda_{kl} + |\hat{X}_{lk}|^2 \right) \right) - \frac{1}{2} \chi^2_{\text{LLG}}$$

(16)

and since the DCT is a linear transformation the mean of the MFCC posterior, which corresponds to the MMSE-MFCC estimation, can be easily computed as

$$\theta_{il}^{\text{MMSE}} = \mu_{\text{CEPS}}^{il} = \sum_{j=1}^{J} T_{ij} \mu_{\text{LLG}}^{il}.$$ 

(17)

In addition to this estimation, the variance of the posterior

$$\Sigma_{\text{LLG}} \approx \sum_{j=1}^{J} T_{ij}^2 \Sigma_{\text{LLG}}^{il}$$

is also available and can be used as a measure of the estimation uncertainty. The resulting estimator defined by Eqs. 7, 8, 15, 16, 17 and 18 can be then computed given the noisy spectrum $Y_{il}$ and estimations of noise and speech variances $\lambda_{D_{kl}}$, $\lambda_{S_{kl}}$ as any other estimator in STFT domain. Furthermore, when adequately implemented, it has a computational cost comparable with twice that of a conventional MFCC feature extraction. For a noiseless environment with $\lambda_{D_{kl}} = 0$, the estima-
tor coincides with the conventional MFCC feature extraction in Eq. 13. If uncertainty is ignored by setting $\lambda_{kl} = 0$, the estimator is equivalent to the conventional MFCC features with a preprocessing using a Wiener filter. Finally since the estimation takes place in the feature domain where the recognizer operates, the estimator variance can be combined with observation uncertainty techniques to further improve the results.

4. Experiments and Results

Table 1: Word error rates (WER) [%] of the AURORA4 tests I. Lowest WER displayed in bold.

<table>
<thead>
<tr>
<th>NoiseType</th>
<th>Clean</th>
<th>Car</th>
<th>Babble</th>
<th>Restaurant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [12]</td>
<td>15.4</td>
<td>49.4</td>
<td>60.6</td>
<td>59.0</td>
</tr>
<tr>
<td>WIENER</td>
<td>14.4</td>
<td>28.7</td>
<td>56.6</td>
<td>58.7</td>
</tr>
<tr>
<td>MMSE-STSA</td>
<td>13.74</td>
<td>31.0</td>
<td>51.0</td>
<td>53.8</td>
</tr>
<tr>
<td>MMSE-LSA</td>
<td>12.9</td>
<td>31.2</td>
<td>50.3</td>
<td>55.0</td>
</tr>
<tr>
<td>MMSE-MFCC</td>
<td>13.3</td>
<td>25.9</td>
<td>46.0</td>
<td>49.8</td>
</tr>
<tr>
<td>MMSE-MFCC+MI</td>
<td>13.2</td>
<td>21.9</td>
<td>42.9</td>
<td>49.0</td>
</tr>
</tbody>
</table>

Table 2: Word error rates (WER) [%] of the AURORA4 tests II. Lowest WER displayed in bold.

<table>
<thead>
<tr>
<th>NoiseType</th>
<th>Street</th>
<th>Airport</th>
<th>Train Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [12]</td>
<td>57.4</td>
<td>61.9</td>
<td>62.0</td>
</tr>
<tr>
<td>WIENER</td>
<td>54.9</td>
<td>59.4</td>
<td>55.2</td>
</tr>
<tr>
<td>MMSE-STSA</td>
<td>45.8</td>
<td>58.2</td>
<td>46.1</td>
</tr>
<tr>
<td>MMSE-LSA</td>
<td>45.1</td>
<td>59.6</td>
<td>43.6</td>
</tr>
<tr>
<td>MMSE-MFCC</td>
<td>44.5</td>
<td>53.7</td>
<td>42.4</td>
</tr>
<tr>
<td>MMSE-MFCC+MI</td>
<td>47.6</td>
<td>45.2</td>
<td>48.6</td>
</tr>
</tbody>
</table>

In order to test the efficiency of the proposed estimator for large vocabulary speech recognition, the AURORA4 database [12] was used together with the Hidden Markov Model Toolkit (HTK). The AURORA4 database corresponds to a noisy version of the Wall-Street Journal (WSJ) database, which has a vocabulary of 5K words. A word internal triphone based ASR system was trained using the routines for HTK available online for the WSJ database [13]. From the AURORA4, the senheisser microphone set with a sampling frequency of 8kHz was used for the tests. Regarding the estimation of noise and speech variances, IMCRA [14], a method for instationary noise variance estimation in STFT domain was employed. The HTK recognition function was also modified to perform speech recognition with uncertain features using the observation uncertainty technique modified imputation [2].

The tests compared the performance of the well known Wiener and Ephraim-Malah estimators (MMSE-STSA, MMSE-LSA) against the proposed MMSE-MFCC estimator with and without using modified imputation (MI). Results displayed in tables 1 and 2 show how the proposed technique outperforms conventional techniques in all scenarios with the exception of a small increase in error rates for clean speech. When combined with observation uncertainty techniques, a majority of the results improve although the performance falls under that of MMSE-STSA and MMSE-LSA estimators for some cases of instationary noise. This is to be expected since observation uncertainty techniques are known to pose problems for large vocabulary recognition. Such problems might be solved with particular modifications which were not used here for maximum generality of the algorithm.

5. Conclusions

We have shown how uncertainty propagation can be used to yield an approximate MMSE-MFCC estimator which outperforms conventional MMSE techniques in STFT domain when used for robust ASR. When combined with observation uncertainty, results also improve in most scenarios. This estimator has also very low computational costs and can be extended to other feature extractions like RASTA-PLPs.

6. References