SPEECH ENHANCEMENT UNDER A COMBINED STOCHASTIC-DETERMINISTIC MODEL

Richard C. Hendriks, Richard Heusdens and Jesper Jensen
Dept. of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands
email: {R.C.Hendriks, R.Heusdens, J.Jensen}@TUDelft.nl

ABSTRACT

Most DFT domain based enhancement methods rely on stochastic models to derive clean speech estimators. In this paper we investigate the use of a deterministic speech model and present an MMSE estimator under a combined stochastic-deterministic speech model. Experimental results show an increase in segmental SNR of 1.18 dB, compared to the use of a stochastic model alone. Furthermore, PESQ evaluations lead to an increase of 0.3 on the MOS scale. Listening tests show a preference for the proposed MMSE estimator under combined stochastic-deterministic speech model.

1. INTRODUCTION

Voice communication systems are often designed for processing of noise free speech. However, speech signals used as an input to these systems are often degraded by acoustical noise. Single microphone speech enhancement methods can be used to reduce the noise level in the noisy signals before they are processed by a voice communication system.

Many such algorithms work in the discrete Fourier transform (DFT) domain, where relatively good quality can be obtained with relatively low computational complexity. Here clean speech DFT coefficients are estimated using a criterion like minimum mean square error (MMSE) [1]. The main focus in DFT domain speech enhancement has been on the derivation of estimators relying completely on a stochastic model for the clean speech DFT coefficients. In practice, speech DFT coefficients have been assumed Gaussian distributed [1], and more recently estimators have been derived which assume Laplacian and Gamma distributions [2].

Although most DFT domain enhancement algorithms rely on stochastic models, it is known that certain speech sounds (e.g., vowels) can be very well modelled with a linear combination of sinusoidal functions with constant frequency and amplitude [3, Ch. 4]. With this signal representation, the sequence of DFT coefficients seen across one particular frequency bin constitutes a completely deterministic time-series. In [4] a maximum likelihood based spectral amplitude estimator was derived under a deterministic speech model. Here, clean speech DFT coefficients are characterized by deterministic, but unknown, amplitude and phase values, while noise DFT coefficients are assumed to follow a zero mean Gaussian pdf.

However, here it was assumed that speech always has a deterministic character, an assumption which is obviously less appropriate for noise-like speech sounds such as /s/, /f/, etc.

We propose in this paper to use a speech model where clean speech DFT coefficients are represented by a mixture between a stochastic and a deterministic model. Using this combined stochastic-deterministic (SD) model, we derive an MMSE clean speech estimator under speech presence uncertainty similar to [1, 4]. By doing so, we exploit the idea that certain speech DFT coefficients can be better modelled using a stochastic representation while others may be better represented with a deterministic one.

2. DETERMINISTIC AND STOCHASTIC SPEECH MODEL

In this section we consider the individual stochastic and deterministic speech models and their MMSE estimators. We assume the noise process to be additive, i.e. $Y(k, i) = X(k, i) + N(k, i)$ with $Y(k, i)$, $X(k, i)$ and $N(k, i)$ the noisy speech, clean speech and noise DFT coefficient for frequency bin $k$ and time frame $i$. Furthermore we assume that $X(k, i)$ and $N(k, i)$ are uncorrelated and that $N(k, i)$ has a zero-mean Gaussian distribution.

2.1. The Stochastic Model

Although different stochastic models have been proposed for speech DFT coefficients [2], we assume here that under the stochastic speech model, speech DFT coefficients have a Gaussian distribution. However, we notice that the presented framework is general and can also be extended to other distributions. Under the Gaussian distribution, the noisy speech DFT coefficients follow a zero-mean complex Gaussian distribution

$$f(Y(k, i)) = \frac{1}{\pi \sigma_Y^2(k, i)} \exp \left\{ -\frac{|Y(k, i)|^2}{\sigma_Y^2(k, i)} \right\},$$

where $\sigma_Y^2(k, i)$ is the variance of the noisy DFT coefficient $Y(k, i)$ which equals the sum of the clean speech and noise variance, that is $\sigma_Y^2(k, i) = \sigma_x^2(k, i) + \sigma_n^2(k, i)$. The MMSE estimator is then given by the conditional mean estimator, known as the Wiener filter,

$$\hat{X}(k, i) = E[X(k, i)|Y(k, i)] = \frac{\xi(k, i)}{1 + \xi(k, i)} Y(k, i).$$

Here, $\xi(k, i) = \frac{\sigma_x^2(k, i)}{\sigma_Y^2(k, i)}$ which is known as the a priori SNR.

2.2. The Deterministic Model

The noisy DFT $Y(k, i)$ under the deterministic speech model is a sum of a deterministic variable $X(k, i)$ and a (zero-mean) Gaussian
Let us therefore assume that our deterministic model for a clean speech DFT coefficient \( k \) is a single complex exponential, that is

\[
X(k, n) = \sum_{m=0}^{K-1} a_p e^{j\phi_p} e^{j\nu_p m} w(m) e^{-j\omega_k m},
\]

where \( w(m) \) is the time sample index, \( a_p \) the amplitude, \( \phi_p \) the phase and \( \nu_p \) the frequency of component \( p \). Under this model, the DFT coefficients at each frequency bin \( k \) can be described by a sum of \( P \) complex exponentials seen across time. However, under the assumption of sufficiently long frame sizes, there will be no more than one dominant exponential, say component \( p \), per frequency bin.

Let us therefore assume that our deterministic model for a clean speech DFT coefficient \( k \) is a single complex exponential, that is

\[
X(k, n) = \sum_{m=0}^{K-1} a_p e^{j\phi_p} e^{j\nu_p m} w(m) e^{-j\omega_k m},
\]

where \( w(m) \), \( m = 0, \ldots, K-1 \) is the analysis window (of length \( K \)) used to define the signal frame, \( M \leq K \) the frame shift and \( \omega_k = \frac{2\pi}{M} \), where \( L \geq K \) is the DFT size. We can write (6) in the form

\[
X(k, n) = \sum_{m=0}^{K-1} a_p e^{j\phi_p} e^{j\nu_p m} w(m) e^{-j\omega_k m},
\]

with \( w(m) \), \( m = 0, \ldots, K-1 \) is the analysis window (of length \( K \)) used to define the signal frame, \( M \leq K \) the frame shift and \( \omega_k = \frac{2\pi}{M} \), where \( L \geq K \) is the DFT size. We can write (6) in the form

\[
X(k, n) = \sum_{m=0}^{K-1} a_p e^{j\phi_p} e^{j\nu_p m} w(m) e^{-j\omega_k m},
\]

where we used the relation in Eq. (6) and where each term is corrected for the phase shift (due to the frame shift).
HARD-SD-U Estimator
The estimator HARD-SD-U follows from Eq. (8) by setting \( f(\beta = D|Y) \) either equal to 1 (deterministic model) or to 0 (stochastic model). This means that
\[
\hat{X} = \begin{cases} 
E[X|Y, \alpha = P, \beta = D] f(\alpha = P|Y, \beta = D) & \text{if det. speech} \\
E[X|Y, \alpha = P, \beta = S] f(\alpha = P|Y, \beta = S) & \text{if sto. speech},
\end{cases}
\]
where the decision between the deterministic and stochastic speech model is made with the following hypothesis test,
\[
H_0 : \quad E[Y(k, i)] = 0 \\
H_1 : \quad E[Y(k, i)] = X(k, i) \text{ and } \text{VAR}[Y(k, i)] = \sigma^2(k, i).
\]
Under the \( H_0 \) hypothesis the stochastic model is chosen and under the \( H_1 \) hypothesis the deterministic model. We decide between \( H_0 \) and \( H_1 \) using the Bayes criterion [7] and compare the likelihood ratio
\[
T = \frac{f(Y|\beta = D)}{f(Y|\beta = S)}
\]
with threshold \( \lambda = \frac{f(\beta = S)}{f(\beta = D)} \).

HARD-SD Estimator
Estimator HARD-SD assumes that \( f(\alpha = P) = 1 \). Therefore,
\[
\hat{X} = E[X|Y] = \int X \sum_{\beta} f(X|Y, \beta) f(\beta|Y) dX.
\]
So that,
\[
E[X|Y] = \begin{cases} 
E[X|Y, \beta = D] & T \geq \lambda \\
E[X|Y, \beta = S] & T < \lambda,
\end{cases}
\]
with \( T \) and \( \lambda \) as given for the HARD-SD-U Estimator.

4. EXPERIMENTAL RESULTS
For objective evaluation segmental SNR is used, which is defined as
\[
\text{SNR}_{seg} = \frac{1}{L} \sum_{i=0}^{L-1} T \left\{ 10 \log_{10} \frac{\| \tilde{x}(i) \|_2^2}{\| x(i) - 2 \tilde{x}(i) \|_2^2} \right\}
\]
where \( x(i) \) and \( \tilde{x}(i) \) denote frame \( i \) of the clean speech signal \( x \) and the enhanced speech signal \( \tilde{x} \), respectively, \( L \) the number of frames within the speech signal in question and \( T(x) = \min\{\max(x, -10), 35\} \), which confines the SNR to a perceptually meaningful range. All objective results are averaged over 10 different speech signals all sampled at 8 kHz. We use frame sizes of 256 samples taken with 50% overlap. For good time resolution in the estimation of (4), the samples \( Y'(k, n) \) are computed from frames with an overlap of 84%. This overlap was chosen based on a trade off, where on one hand a small overlap is desirable, to better satisfy the assumption made in Section 2. On the other hand, a large overlap is necessary when using multiple samples in (7), i.e. \( n_1, n_2 > 0 \), because approximation of (4) by (7) is only valid over relatively short time intervals. In all experiments, noise statistics are measured during silence regions preceding speech activity.

Furthermore, we chose \( n_1 = 2 \) and \( n_2 = 2 \) and the prior probabilities as \( f(\beta = D) = 0.02 \) and \( f(\alpha = P|\beta = D) = 0.01 \) and \( f(\alpha = P|\beta = S) = 0.2 \), based on performance in terms of \( \text{SNR}_{seg} \). For the proposed methods the decision directed [1] approach is used to estimate the a priori SNR \( \xi(k, i) \). We chose \( \alpha = 0.98 \) based on initial listening experiments. For comparison we use the MMSE estimator under Gaussian distribution (Eq.(2)), where \( \xi(k, i) \) is estimated with the decision directed approach with \( \alpha = 0.97 \) as proposed in [1] for the Wiener filter.

4.1. Objective Results
In Fig. 2 we evaluate the performance of the presented algorithms SOFT-SD-U, HARD-SD-U and HARD-SD and compare that with the use of a stochastic model alone (Wiener filter), for speech signals degraded by white noise at an SNR in the range from 0 dB to 20 dB. Over the whole range of input SNRs the proposed methods improve the performance compared to the use of a stochastic model alone. In terms of segmental SNR, the performance improvement of HARD-SD over the use of a stochastic model alone is 0.14 dB. Incorporating speech presence uncertainty, i.e. HARD-SD-U over HARD-SD leads to another 0.44 dB improvement. Incorporating a soft decision between the stochastic and the deterministic model, i.e. SOFT-SD-U over HARD-SD-U, leads to an additional 0.60 dB improvement. Altogether, the improvement of SOFT-SD-U over the use of a stochastic model alone is 1.18 dB. In Fig. 3 objective results are shown for signals degraded by F16-fighter cockpit noise. The figure shows similar performance as for the white noise case.

4.2. Subjective Evaluation
We use an extension of the perceptual evaluation of speech quality (PESQ) measure [9], to get a first indication of the subjective quality
of the proposed SOFT-SD-U algorithm. In Fig. 4a and 4b we compare the PESQ scores of SOFT-SD-U with a system where a stochastic model alone is used, for speech signals degraded by white noise and F16-fighter cockpit noise, respectively. Fig. 4a and 4b show that for signals degraded by white noise and F16-fighter cockpit noise respectively, at high input SNRs a PESQ improvement of 0.3, while at low input SNRs the performance difference appears to vanish.

For further subjective evaluation, two OAB listening tests were performed with seven participants, the authors not included. Here, O is the original signal and A and B are two enhanced signals. The listeners were presented first the original signal followed by two different enhanced versions. Each series was repeated 4 times with the enhanced versions played in random order. We used 4 different speech signals from the Timit database, two female speakers and two male speakers, degraded by white noise at SNRs of 5 and 15 dB.

First, we evaluated the perceptual performance of using both the speech presence estimator and the soft decision SD speech model (SOFT-SD-U) compared to the use of a stochastic speech model alone. The average relative preference for the SOFT-SD-U was 88% and 83% at 5 and 15 dB SNR, respectively.

In the second listening test, we concentrated on the perceptual impact of using a combined SD speech model instead of a stochastic speech model alone. To do so we implemented a Wiener filter combined with a speech presence uncertainty estimator and compared that with our proposed method SOFT-SD-U, that also incorporates speech presence uncertainty. The average relative preference for the SOFT-SD-U method over the Wiener filter with speech presence uncertainty was 63% and 59% at 5 and 15 dB SNR, respectively.

Comments given by the participants on the tested signals was that the proposed SOFT-SD-U method leads to less suppressed and less muffled speech sounds. However, this leads to more dynamics in the enhanced signal, which was not appreciated by all listeners. At 15 dB, SOFT-SD-U was said to reduce the amount of echo like speech distortions that are present using the stochastic model alone.

5. CONCLUSIONS

We presented an MMSE estimator under a combined stochastic-deterministic speech model. Under the deterministic speech model, clean speech DFT coefficients are modelled as a complex exponential across time with constant amplitude. Under the stochastic speech model the speech DFT coefficients are assumed to be Gaussian distributed. The presented method is general and can be extended to other distributions. With objective experiments and PESQ evaluation it was shown that the proposed MMSE estimator leads to improvements over the use of a stochastic speech model alone. Moreover, listening experiments demonstrated a preference for the proposed method.

6. REFERENCES