SPEECH ENHANCEMENT BASED ON RAYLEIGH MIXTURE MODELING OF SPEECH SPECTRAL AMPLITUDE DISTRIBUTIONS

J.S. Erkelens*, J. Jensen†, and R. Heusdens*

* Delft University of Technology, Department of Mediamatics, Mekelweg 4, 2628 CD Delft, The Netherlands
† Oticon A/S, Kongebakken 9, 2765 Smørum, Denmark

ABSTRACT

DFT-based speech enhancement algorithms typically rely on a statistical model of the spectral amplitudes of the noise-free speech signal. It has been shown in the literature recently that the speech spectral amplitude distributions, conditional on estimated a priori SNR, may differ significantly from the traditional Gaussian model and are better described by super-Gaussian probability density functions. We show that these conditional distributions can be accurately approximated by a mixture of Rayleigh distributions. The MMSE amplitude estimators based on Rayleigh Mixture Models perform at least as well as the estimators based on super-Gaussian models. Furthermore, the proposed Rayleigh Mixture Models allow for derivation of closed-form estimators minimizing other perceptually relevant distortion measures, which may be difficult for other models.

1. INTRODUCTION

The traditional assumption for speech enhancement in the DFT domain is that the distribution of the complex speech DFT coefficients is Gaussian [1]–[3]. Consequently, the spectral amplitude distribution is modeled by a Rayleigh distribution. Recently, super-Gaussian models of the DFT coefficients have received quite some attention, because they lead to estimators with better performance than those based on a Gaussian model. Martin [4] derived complex-DFT estimators for Laplacian and Gamma speech priors, and Lotter and Vary [5] proposed a Maximum A Posteriori (MAP) amplitude estimator for a generalized Gamma amplitude distribution. MMSE estimators for the amplitudes, assuming a one-sided generalized Gamma distribution, were treated in [6] and [7].

In this paper we propose to model the distributions of speech DFT amplitudes by Rayleigh Mixture Models (RMMs). RMMs have some important advantages over existing speech models. They offer more accurate fits to the speech spectral amplitude distributions, conditional on estimated a priori SNR, may differ significantly from the traditional Gaussian model and are better described by super-Gaussian probability density functions. We show that these conditional distributions can be accurately approximated by a mixture of Rayleigh distributions. The MMSE amplitude estimators based on Rayleigh Mixture Models perform at least as well as the estimators based on super-Gaussian models. Furthermore, the proposed Rayleigh Mixture Models allow for derivation of closed-form estimators minimizing other perceptually relevant distortion measures, which may be difficult for other models.

2. MMSE SPECTRAL ESTIMATION

2.1 Signal model and assumptions

We consider an additive-noise signal model of the form

\[ X(k, m) = S(k, m) + D(k, m), \]

where \( X(k, m), S(k, m), \) and \( D(k, m) \) are complex-valued random variables representing the short-time DFT coefficients obtained at frequency index \( k \) in signal frame \( m \) from the noisy speech, clean speech and noise process, respectively. Applying the standard assumption that \( S(k, m) \) and \( D(k, m) \) are statistically independent across time and frequency as well as from each other, leads to expressions for the resulting estimators that are independent of time and frequency. For ease of notation we therefore drop the time and/or frequency index when this does not cause confusion. We use capitals for random variables and the corresponding lower-case letters for their realizations. The speech amplitude is \( A = |S| \), and the noisy amplitude is \( R = |X| \). The noise DFT coefficients \( D \) are assumed to follow a complex Gaussian distribution with variance \( \lambda_D \).

2.2 p-th Order amplitude estimators

For given noise spectral variance \( \lambda_D \) and given speech spectral variance \( \lambda_S \), the MMSE estimator of some power \( p \) of the speech amplitude is (see, e.g., [1]):

\[
\hat{A}^p = E[A^p|R] = \frac{\int_0^{\lambda_S} da f_A(a)|a|^p J_0\left(\frac{2\lambda_D a}{\lambda_S}\right) da}{\int_0^{\lambda_S} da f_A(a) J_0\left(\frac{2\lambda_D a}{\lambda_S}\right) da},
\]

\[ (1) \]

where \( f_A(a) \) is the probability density function of \( A \), which depends on \( \lambda_S \), and \( J_0(\cdot) \) is the zeroth order Bessel function of the fist kind. \( \hat{A}^p \) is called the \( p \)-th order amplitude estimator. In practice, the spectral variances \( \lambda_S \) and \( \lambda_D \) are unknown and have to be estimated. This will affect the optimality of the estimators. We will take into account, to some extent, the influence of the speech spectral variance estimation, by matching our model of \( f_A(a) \) to measured histograms that are conditional on a high value of estimated \( \lambda_S \). In the following, we assume that the noise spectral variance can be estimated accurately during speech pauses for stationary noise or by using approaches based on minimum-statistics [8][9], for example, for non-stationary noise.

2.2.1 Generalized Gamma distribution

Recently, the clean-amplitude distribution \( f_A(a) \) in (1) has been modeled using the generalized Gamma distribution [5,
6, 7]. This distribution is given by
\[ f_X(a) = \frac{2a^2}{\Gamma(\nu)} \exp(-\beta a^2), \quad a \geq 0, \tag{2} \]
with the constraints on the parameters \( \gamma > 0, \nu > 0 \). We will consider the cases \( \gamma = 1 \) and \( \gamma = 2 \). Because \( E\{A^2\} \) equals \( \lambda_S \) by definition, \( \beta \) is related to \( \gamma, \nu \) and \( \lambda_S \). For \( \gamma = 1 \) we have \( \beta^2 = \nu(v+1)/\lambda_S \), and \( \beta = \sqrt{\nu}/\lambda_S \) for \( \gamma = 2 \). For \( \gamma = 2 \), \( \nu = 1 \), the Rayleigh distribution appears as a special case.

2.2.2 Rayleigh Mixture Model
As an alternative to the generalized Gamma model, we propose a Rayleigh Mixture Model. If the complex speech DFT coefficients are modeled by a Gaussian Mixture Model, then the amplitude distribution is a sum of Rayleigh distributions:
\[ f_X(a) = \sum_{j=1}^{J} C_j \frac{2a}{\lambda_j} \exp\left(-\frac{a^2}{\lambda_j}\right), \tag{3} \]
where \( J \) is the number of components and the \( C_j \) are positive weighting factors that satisfy \( \sum C_j = 1 \). The \( \lambda_j \) are the variances of the individual components; they satisfy \( \sum C_j \lambda_j = \lambda_S \).

2.3 A priori SNR estimation
Speech amplitude estimators are usually written in terms of gain functions, e.g., \( \hat{A} = G(\xi, \zeta)R \). These gain functions depend on a priori SNR \( \xi \), defined as \( \xi = \lambda_S/\lambda_D \), and a posteriori SNR \( \zeta \), defined as \( \zeta = R^2/\lambda_D \). We will use the decision-directed approach [1] to estimate a priori SNR, with a bias correction [10]:
\[ \hat{\xi}_k(m) = \max(\alpha \frac{\lambda_k^2(m-1)}{\lambda_D(k,m-1)} + (1-\alpha)[\hat{\xi}_k(m) - 1], \xi_{\text{min}}). \tag{4} \]
Note that in the first term, the second order amplitude estimator is used, instead of the square of the first order amplitude estimator, which was the original definition [1]. The second order amplitude estimator used in (4) will be based on the generalized Gamma distribution (2), with either \( \gamma = 1 \) or \( \gamma = 2 \). We have observed that this new a priori SNR estimator (4) leads to less musicality than the old definition, for parameter settings \((\nu, \alpha)\) with the same speech quality versus noise reduction trade-off [11].

3. RAYLEIGH MIXTURE MODELING OF CONDITIONAL SPEECH AMPLITUDE DISTRIBUTIONS
It has been shown in several papers [4]–[7] that better noise suppression performance can be achieved by abandoning the Gaussian speech model. There may be several reasons for the suboptimality of the Gaussian model. Often, the normal distribution of DFT coefficients is motivated by the central limit theorem. For speech DFT coefficients, the central limit theorem may not be applicable, because of the long span of correlation which can be larger than the frame lengths [4][5]. Speech is also non-stationary, causing many time frames to contain non-identically distributed samples [6]. Furthermore, gain functions are derived for known a priori SNR. In practice, a priori SNR has to be estimated. This means that the optimal statistical model for enhancement may differ from the true underlying speech distribution, and should be adapted to the a priori SNR estimator used [10]–[13].

3.1 Measured amplitude distributions
Following an idea of Martin [4], Lotter and Vary [5] have attempted to measure the distribution of amplitudes of speech DFT coefficients. For this purpose, a speech database is processed in a standard DFT-based enhancement algorithm, and coefficients are collected from those frequency bins for which the estimated a priori SNR is within a narrow range of high values. We performed a similar experiment. Figure 1(a) shows a histogram of one million of such amplitudes from TIMIT, normalized such that the second moment equals one, i.e., \( \lambda_S = 1 \), where the overbar indicates the sample mean. Also shown are maximum-likelihood fits of a Rayleigh distribution and generalized Gamma distributions. Clearly the measured amplitude distribution does not follow the Rayleigh model, while the generalized Gamma models fit better. Amplitude estimators based on generalized Gamma distributions improve speech enhancement performance over those based on a Gaussian speech model [5]–[7]. Figure 1(b) shows a maximum-likelihood fit (see section 4.2.1) of the proposed RMM (3) to the histogram, using \( J = 7 \) components. Clearly, the RMM model offers a much better fit to the histogram. The experiments of Section 5 show that the resulting estimators also perform very well in a speech enhancement context. However, online adaptation to speech characteristics would be easier for the generalized Gamma models, because of the smaller number of parameters.

3.2 Discussion
Ephraim and Cohen [14] have shown that the Gaussian speech model and other models are not necessarily contradictory. If the spectral variance \( \lambda_S \) is treated as a random variable with pdf \( f(\lambda_S) \), then the joint distribution of real and imaginary parts of the corresponding DFT coefficient is given by
\[ f(s_R, s_I) = \int f(s_R, s_I | \lambda_S)f(\lambda_S)d\lambda_S, \]
where \( s_R \) and \( s_I \) are the real and imaginary parts of a clean speech DFT coefficient, respectively. If \( f(\lambda_S) \) is a Gaussian distribution, then \( f(s_R, s_I | \lambda_S) \) is a continuous mixture of Gaussian distributions, which can take many different forms depending on \( f(\lambda_S) \). For example, if \( f(\lambda_S) \) is assumed to be exponential, then the pdf of the real and imaginary parts each follow a Laplace pdf, as it was assumed in [4].
The distributions faced in practice are conditional on estimated a priori SNR and are given by
\[ f(s_R,s_I|\hat{\lambda}_S) = \frac{1}{\hat{S}_\lambda} f(s_R,s_I|\hat{\lambda}_S) f(\hat{\lambda}_S)d\hat{\lambda}_S, \]
where \( \hat{\lambda}_S \) is usually done in the derivation of estimators. If \( \hat{\lambda}_S \) is only an estimate of the true spectral variance \( \lambda_S \), it may contain less information about \( s_R \) and \( s_I \) than \( \lambda_S \) itself. The second term in the decision directed estimator (4), \( (1-\alpha)|f^2(m)/\lambda_D - 1| \), depends on the noisy amplitude and therefore contains some information about the clean \( s_R \) and \( s_I \). However, the weighting factor \( (1-\alpha) \) of this term is generally small (0.02 is a typical value). We therefore expect the following approximation to be reasonable:
\[ f(s_R,s_I|\hat{\lambda}_S) \approx \int_0^\infty f(s_R,s_I|\hat{\lambda}_S) f(\hat{\lambda}_S)d\hat{\lambda}_S. \]
Note that any dependency that may exist between the real and imaginary parts of the current time frame, \( s_R(m) \) or \( s_I(m) \), and the estimated (second order) amplitude of the previous time frame, \( \hat{\lambda}_S(m-1) \), is also neglected given \( \hat{\lambda}_S \) as is usually done in the derivation of estimators. If \( f(s_R,s_I|\hat{\lambda}_S) \) is Gaussian, then (5) expresses \( f(s_R,s_I|\hat{\lambda}_S) \) as a continuous mixture of Gaussians. The corresponding amplitude distribution is a continuous mixture of Rayleigh distributions. We propose to model such amplitude distributions by RMMs (3). That model is used in (1) to obtain estimators that take into account statistics of the speech and the particular a priori SNR estimator used. Note that we do not really rely on (5), because the RMM model can accurately match the histograms with a sufficiently large number of components, regardless of whether (5) is accurate or not. As was illustrated in Figure 1(b), only a small number of components suffices in practice.

4. AMPLITUDE ESTIMATORS

4.1 Generalized Gamma distribution
A MAP amplitude estimator for the model (2) for \( \gamma = 1 \) was derived in [5], while MMSE amplitude estimators for the classes \( \gamma = 1 \) and \( \gamma = 2 \) have been studied in [6] and [7]. For \( \gamma = 2 \), the expressions are exact, while approximations have to be made for \( \gamma = 1 \). The maximum achievable performance for both classes is about the same. Because of lack of space, we show only the expressions for the estimators of the \( \gamma = 2 \) class, which contain the Gaussian speech model as a special case for \( \gamma = 1 \). The MMSE amplitude estimator is given by
\[
\hat{\lambda}_S^{(2)} = \frac{\Gamma(v + 0.5)}{\Gamma(v)} \sqrt{\frac{\xi}{\xi(v + \xi)}} \Gamma_1 \left( v + 0.5; 1; \frac{\xi v}{\xi + v} \right) R, 
\]
where \( \Gamma_1(a; b; x) \) is a confluent hypergeometric function [15, 13.1.2]. The superscript \( (2) \) indicates that \( \gamma = 2 \). The corresponding second order amplitude estimator is given by
\[
\hat{\lambda}_S^{(2)} = \frac{v \xi}{\xi(v + \xi)} \Gamma_1 \left( v + 1; 1; \frac{\xi v}{\xi + v} \right) R. 
\]
-19 dB. We use all 30 clean sentences of the NOIZEUS database [16]. Noisy signals were generated by adding white and nearly stationary car noise from the NoiseX-92 database [17] to the clean signals, at 5 and 15 dB overall SNR. The noise and speech are limited to telephone bandwidth (300-3400 Hz). The noise variance was estimated from 0.64 seconds of noise only preceding speech activity. Objective quality was measured in two different ways. We measure mean-square error (MSE), because it is what MMSE estimators should minimize on the average. We compute MSE as

\[
MSE = \frac{1}{M} \sum_{m=1}^{M} \sum_{k} \{a(k,m) - \hat{a}(k,m)\}^2,
\]

where \(a(k,m)\) and \(\hat{a}(k,m)\) are the clean speech spectral amplitude and the estimated amplitude of frequency bin \(k\) and time frame \(m\), respectively, and \(M\) is the number of frames containing speech in a sentence. To exclude silence intervals, frames with a clean energy more than 40 dB below the maximum clean frame energy of a speech sentence are not taken into account. All results at a given SNR are averages over all test sentences. Furthermore, to quantify the speech distortion versus noise reduction trade-off, we also measure separately Segmental Speech Quality (SQ) and Noise Reduction (NR) as in [5], and plot these quantities against each other while varying \(\nu\). The enhanced speech \(\hat{s}(n)\) can be written as

\[
\hat{s}(n) = \hat{s}(n) + \hat{d}(n),
\]

where \(\hat{s}(n)\) and \(\hat{d}(n)\) result from applying the gain functions to the clean speech and noise DFT coefficients separately, and transforming back to the time domain. We define Segmental Speech Quality as

\[
SQ = \frac{1}{M} \sum_{m=1}^{M} 10 \log_{10} \left( \frac{||s_{m}||^2}{||s_{m} - \hat{s}_{m}||^2} \right),
\]

where \(s_{m}\) and \(\hat{s}_{m}\) denote time frame \(m\) of the signals \(s(n)\) and \(\hat{s}(n)\), respectively. The operator \(||\cdot||^2\) computes the energy of a time frame. Segmental Noise Reduction is defined as

\[
NR = \frac{1}{M} \sum_{m=1}^{M} 10 \log_{10} \left( \frac{||d_{m}||^2}{||q_{m}||^2} \right).
\]

Strong suppression leads to low SQ and high NR, while the opposite happens for weak suppression.

5.2 Performance evaluation

We will compare amplitude estimators for the generalized Gamma model with those of RMM models, while varying the parameter \(\nu\). A priori SNR estimation with (4) was always based on the generalized Gamma model. The parameters of the RMM models are found from the corresponding histograms, as was outlined in Sections 3.1 and 4.2.1. Figure 2 shows the results for the \(\gamma = 1\) case. The dash-dotted lines result for white noise when the generalized-Gamma amplitude estimators are used for reconstruction, while the solid curves are for RMM amplitude estimators with \(J = 7\) components. The dotted and dashed lines are the corresponding results for car noise. The value of \(\nu\) is limited to values larger than 0.5 for \(\gamma = 1\), because of an approximation that is used in the derivation of the estimators [7]. The crosses and pluses result for the RMM amplitude estimator, for white and car noise respectively, when an exponential smoother is used for a priori SNR estimation (corresponding to \(\nu \to \infty\)). The upper two panels show that minimum achievable MSE is lower for the RMM amplitude estimators. Furthermore, the RMM estimators are much less sensitive to the value of \(\nu\). The main reason for this behavior is that the RMM models have been adapted to some extent to the a priori SNR estimator used, because the parameters are found from measured data that depend on it (see sections 3.1 and 4.2.1). It is clear that using an exponential smoother for a priori SNR estimation is not optimal. The lower two panels show SQ versus NR when \(\nu\) is varied over the same range as for the upper two panels. The value of \(\nu\) increases when going from the right to the left along the curves. More noise reduction is possible for the generalized Gamma amplitude estimators, but the maximum achievable speech quality is higher for the RMM amplitude.
estimators. Similar trends are seen for car noise.

Figure 3 shows the results for the $\gamma = 2$ case. The maximum achievable performance is about the same as for the $\gamma = 1$ case, although the results are much more sensitive to the value of the $\nu$-parameter. The RMM amplitude estimators perform about the same on the $a$ priori SNR estimators of both cases.

5.2.1 Informal listening

For increasing values of $\nu$ the enhanced sound becomes more reverberant but the musicality decreases, especially for the amplitude estimators of the $\gamma = 2$ class. For the RMM amplitude estimators the $\nu$-dependency is the weakest, although these effects are clearly noticeable for $\nu \to \infty$. For the lowest values of $\nu$, all estimators sound very similar.

5.3 Discussion

Amplitude and $a$ priori SNR estimation for the generalized Gamma models is based on one and the same prior speech distribution (i.e., (2) with the same values of $\gamma$ and $\nu$). This does not necessarily lead to optimal results. To a given $a$ priori SNR estimator corresponds a certain measured histogram of spectral amplitudes. This histogram depends on the unknown dynamical and statistical properties of this particular $a$ priori SNR estimator. There is no reason why the parametric amplitude distribution used in calculating the $a$ priori SNR estimates should fit accurately to its corresponding measured amplitude histogram. In fact, we have seen that an RMM model can fit much better to histograms found with the generalized gamma $a$ priori SNR estimator. We have not investigated whether using different $\gamma$ and/or $\nu$ values for the amplitude and $a$ priori SNR estimation tasks leads to significant improvements for the generalized Gamma models.

Many estimators found in literature that are based on parametric models of the speech prior distribution, including the ones presented here, are implicitly assuming that the conditional distribution $f_x(a|\lambda_S)$ has the same shape (except for a variance scaling) for all values of $\hat{\lambda}_S$. This may not be an accurate assumption for all SNRs, because the properties of the $a$ priori SNR estimator depend on the SNR. A data-driven approach has been proposed in [12] to deal with this problem.

6. SUMMARY AND CONCLUSIONS

In this paper we have proposed Rayleigh Mixture Models to describe measured speech amplitude distributions in the context of speech enhancement. We have shown that the resulting amplitude estimators can compete with state-of-the-art estimators. Furthermore, analytical derivation of estimators for meaningful distortion measures is relatively simple.

REFERENCES


