A New Statistical Excitation Mapping for Enhancement of Throat Microphone Recordings

M.A. Tuğtekin Turan and Engin Erzin
Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Istanbul, Turkey
mturan,eerzin@ku.edu.tr

Abstract

In this paper we investigate a new statistical excitation mapping technique to enhance throat-microphone speech using joint analysis of throat- and acoustic-microphone recordings. In a recent study we employed source-filter decomposition to enhance spectral envelope of the throat-microphone recordings. In the source-filter decomposition framework we observed that the spectral envelope difference of the excitation signals of throat- and acoustic-microphone recordings is an important source of the degradation in the throat-microphone voice quality. In this study we model spectral envelope difference of the excitation signals as a spectral tilt vector, and we propose a new phone-dependent GMM-based spectral tilt mapping scheme to enhance throat excitation signal. Experiments are performed to evaluate the proposed excitation mapping scheme in comparison with the state-of-the-art throat-microphone speech enhancement techniques using both objective and subjective evaluations. Objective evaluations are performed with the wideband perceptual evaluation of speech quality (ITU-PESQ) metric. Subjective evaluations are performed with the A/B pair comparison listening test. Both objective and subjective evaluations yield that the proposed statistical excitation mapping consistently delivers higher improvements than the statistical mapping of the spectral envelope to enhance the throat-microphone recordings.

Index Terms: throat-microphone, speech enhancement, spectral envelope mapping, excitation mapping, GMM mapping.

1. Introduction

Throat microphone (TM) is a skin-attached piezo-electric transducer that captures speech signal in the form of vibrations through the neck. Due to its direct connection with the skin, it is robust under noisy environments, yet it suffers from the perceived speech quality and intelligibility. Similarly, the non-audible murmur (NAM) microphone has been introduced by Nakajima et al. [1] to detect body-conducted speech. The NAM speech has been widely studied as a silent speech interface. One advantage of TM recordings over NAM speech is that the TM recordings can capture the fundamental frequency together with a low-pass version of the vocal track resonance structure. TM recordings deliver attractive signal representations for robust speech processing applications under adverse noise conditions, such as airplane, motorcycle, military field, factory or street crowd environments.

The use of non-acoustic sensors in multi-sensory speech processing has been studied for speech enhancement, robust speech modeling and improved speech recognition [2, 3, 4, 5, 6, 7]. Speech enhancement of non-acoustic sensor recordings employs techniques used for voice conversion [8, 9] and artificial bandwidth extension [10, 11] to improve naturalness and intelligibility of the speech signal. One widely used framework for enhancement of the non-acoustic sensor recordings is the source-filter decomposition, which breaks down the problem into two, namely the enhancement of the excitation (source) and the enhancement of the spectral envelope (filter).

Enhancement of the spectral envelope has been both studied for the speech conversion and the artificial bandwidth extension problems. Stylianou et al. [8] presented one of the early works on continuous probabilistic mapping of the spectral envelope for the voice conversion problem. Graciarena et al. used a similar mapping, the probabilistic optimum filter (POF) mapping, for the estimation of clean acoustic speech features from combined throat and acoustic microphone recordings [2]. Later Toda et al. [9] improved the continuous probabilistic mapping by incorporating not only static but also dynamic feature statistics for the estimation of a spectral parameter trajectory. In [6], we develop a framework to define a temporal correlation model between simultaneously recorded throat- and acoustic-microphone speech. The resulting temporal correlation model is then employed to estimate acoustic features, which are spectrally richer than throat features, from throat features through linear prediction analysis. Spectral envelope mapping techniques have been also studied extensively for the artificial bandwidth extension of telephone speech [10, 11].

Enhancement of the excitation has been studied on domain specific problems and not as widely as the enhancement of spectral envelope. Recently, conversion methodologies from NAM to acoustic and whispered speech have been developed to improve voice quality and intelligibility of NAM speech [7]. In [7] spectral and excitation features of acoustic speech are estimated from the spectral feature of NAM. Since NAM lacks fundamental frequency information, a mixed excitation signal is estimated based on the estimated fundamental frequency and aperiodicity information from NAM. The converted speech reported to suffer from unnatural prosody. In another study [12], the transfer characteristics of bone-conducted and acoustic-microphone speech signals are modeled as dependent sources, and an equalizer, which is trained using simultaneously recorded acoustic and bone-conducted microphone speech, has been investigated to enhance bone-conducted speech. In the bandwidth extension framework, the extension of the excitation signal has been performed by modulation, which attains spectral continuation and a matching harmonic structure of the baseband [10].

In our recent work [13], we developed a speaker- and phone-dependent probabilistic mapping from TM speech spectra to acoustic-microphone (AM) speech spectra for enhancement of the TM speech. The probabilistic mapping is trained using simultaneously recorded AM and TM speech. In this paper we investigate a new statistical excitation mapping technique...
to enhance throat-microphone speech. In the source-filter decomposition framework we observed that the spectral envelope difference of the excitation signals of TM and AM recordings is an important source of the degradation for the TM voice quality. In this paper we model spectral envelope difference of the excitation signals as a spectral tilt vector, and we propose a new phone-dependent probabilistic spectral tilt mapping scheme to enhance throat excitation signal. The organization of this paper as follows: In section 2 we introduce the proposed throat-microphone speech enhancement system. Section 3 discusses the experimental evaluations and results. Finally, section 4 includes the discussions and future research directions.

2. Enhancement System

2.1. GMM based probabilistic mapping

The Gaussian mixture model (GMM) estimator of [8, 14] is a soft mapping from observable source \( X \) to hidden source \( Y \) with an optimal linear transformation in the minimum mean square error (MMSE) sense. This mapping can be formulated as with an optimal linear transformation in the minimum mean square error (MMSE) sense. This mapping can be formulated as

\[
\hat{y}_k = \sum_{l=1}^{L} p(\gamma_l | x_k) [\mu_{y,l} + C_{yx,l}(C_{xx,l})^{-1} (x_k - \mu_{x,l})],
\]

where \( \gamma_l \) is the \( l \)-th Gaussian mixture and \( L \) represents the total number of Gaussian mixtures. The vectors \( \mu_{x,l} \) and \( \mu_{y,l} \) are respectively the centroids for the \( l \)-th Gaussian for sources \( X \) and \( Y \), \( C_{xx,l} \) is the covariance matrix of source \( X \) in the \( l \)-th Gaussian, and \( C_{yx,l} \) is the cross-covariance matrix of sources \( X \) and \( Y \) for the \( l \)-th Gaussian mixture. The probability of the \( l \)-th Gaussian mixture given the observation \( x_k \) is defined as the normalized Gaussian pdf as,

\[
p(\gamma_l | x_k) = \frac{N(x_k; \mu_{x,l}, C_{xx,l})}{\sum_{m=1}^{L} N(x_k; \mu_{x,m}, C_{xx,m}).}
\]

2.2. Enhancement framework

Let us consider having two simultaneously recorded TM and AM speech, which are represented as \( s_T[n] \) and \( s_A[n] \), respectively. Source-filter decomposition through the linear prediction filter model of speech can be defined as,

\[
S_T(z) = \frac{1}{W_T(z)} R_T(z), \quad S_A(z) = \frac{1}{W_A(z)} R_A(z),
\]

where \( W_T(z) \) and \( W_A(z) \) are the inverse linear prediction filters, and \( R_T(z) \) and \( R_A(z) \) are the source excitation spectra for the TM and AM speech, respectively. Then we can define the TM speech enhancement problem as finding two mappings, the first one from TM spectra to AM spectra, and the second one from TM excitation to AM excitation,

\[
\hat{W}_A(z) = \varphi_D(R_T(z) | A_{TA}^{x}), \quad \hat{R}_A(z) = \varphi_D(R_T(z) | A_{TA}^{x}).
\]

where \( A_{TA}^{x} \) and \( A_{TA}^{y} \) are general correlation models of TM and AM spectral envelopes and excitation. These joint correlation models can be extracted using a simultaneously recorded training database. Replacing the TM speech spectra and excitation with the estimates,

\[
\hat{S}_A(z) = \frac{1}{W_A(z)} \hat{R}_A(z),
\]

is expected to enhance the perceived quality of the TM speech.

2.3. Spectral envelope enhancement

In this study, the line spectrum frequency (LSF) feature vector representation of the linear prediction filter is used to model spectral envelope. The TM and AM spectral representations are extracted as 16th order linear prediction filters over 10 ms time frames. We define the elements of this representation at time frame \( k \) as column vectors \( x_k \) and \( y_k \), respectively representing the TM spectral envelope as an observable source \( X^e \) and AM spectral envelope as a hidden source \( Y^e \). Then, a phone-dependent mapping for the spectral envelope enhancement of TM recordings is defined as following,

\[
y_{ke} = \sum_{c=1}^{C} p(\gamma^c_k | x_k) [\mu^c_{y,k} + C^c_{yx,k}(C^c_{xx,k})^{-1} (x_k - \mu^c_{x,k})].
\]

where \( \gamma^c \) is the given phone, and \( L \) is the total number of Gaussian mixtures for the phone class \( c \).

2.4. Excitation enhancement

In the source-filter decomposition framework we observed that the TM and AM recordings exhibit significant differences at the excitation signal spectra, which appears to be an important source of the degradation in the TM voice quality. In this study we model spectral envelope difference of the excitation signals as a spectral tilt vector, and we propose a new phone-dependent GMM-based spectral tilt mapping scheme to enhance TM excitation.

Let us first define a triangular filterbank, which will help us to compute the average spectrum around a sequence of center frequencies,

\[
w_{mb}(n) = \begin{cases} 
0 & n < f_{b-1} \text{ or } n > f_{b+1} \\
\frac{n-f_{b-1}}{f_b-f_{b-1}} & f_{b-1} \leq n \leq f_b \\
\frac{f_b-n}{f_{b+1}-f_b} & f_b < n \leq f_{b+1} 
\end{cases}
\]

where \( f_b \) is the \( b \)-th center frequency index and \( w_{mb} \) is the \( b \)-th triangular filter. We take number of bands as \( B = 8 \) and let \( b = 0, 1, \ldots, B+1 \), where \( f_0 = 0 \) and \( f_{B+1} = N \) are taken as boundary frequency indexes. Then the average spectrum energy of the AM excitation signal is computed for frequency band \( b \) as,

\[
E_A(b) = \log \left( \frac{1}{N-1} \sum_{n=1}^{N-1} w_{mb}(n) | R_A(n) |^2 \right) \text{ for } b = 1, \ldots, B,
\]

where \( R_A \) is the \( 2N \)-point DFT of the excitation signal, and \( B \) is the total number of frequency bands. Similarly the average spectrum energies of the TM excitation can be computed and represented as \( E_T \).

We can now define the spectral tilt vector between the TM and AM excitation signals as,

\[
D(b) = E_A(b) - E_T(b) \text{ for } b = 1, \ldots, B.
\]

The spectral tilt vector is considered as the representation of the hidden source \( Y^e \) for the probabilistic excitation mapping.
define the spectral tilt vector at time frame \( k \) as column vector \( \hat{y}_T[k] \) representing the hidden source \( Y_T \). The observable source of the spectral envelope mapping, which is the 10th order LSF feature vector \( x_T[k] \) of TM speech can be considered as a valuable observation also for the probabilistic excitation mapping. However we also consider excitation spectrum of the TM speech to be valuable. Hence we compute a cepstral feature vector representing the TM excitation spectrum as,

\[
c_T(n) = \sum_{b=1}^{B} E_T(b) \cos(\pi n(b - 1/2)/B),
\]

for \( n = 1, 2, \ldots, B - 1 \). We form the observable source vector of the excitation mapping as \( x_T = [x_0^T \ c_T^T]^T \), where \( c_T \) representing the cepstral column feature vector at frame \( k \). Then, a phone-dependent mapping for the excitation enhancement of TM recordings is defined similarly as in (8) to estimate \( \hat{y}_T[e] \), or equivalently the spectral tilt vector \( \hat{D} \).

Finally, the enhanced excitation spectrum can be estimated by tilting the TM excitation spectrum as following,

\[
\hat{R}_A^D(n) = \sum_{b=0}^{B+1} \sum_{b=0}^{N-1} M_{e} \left( \cos(\pi n(b - 1/2)/B) \right),
\]

where boundary spectral tilt values are taken as \( D_0 = \hat{D}_1 \) and \( D_{B+1} = \hat{D}_B \).

Note that in processing of the excitation signals, a 2048-point DFT is used over 20 ms hamming windowed excitation signals with a frame shift of 10 ms. The enhanced excitation signal is reconstructed from the \( \hat{R}_A^D \) spectrum with inverse DFT and overlap-and-add schemes.

3. Experimental Evaluations

We perform experiments on two synchronous TM and AM databases. The first one is a male speaker (M1) that we used in [6, 13], and the second one is a new male speaker (M2) with a new IASUS GP3 throat microphone system. Each speaker utters 799 sentences, which are recorded synchronously at 16-kHz sampling rate. We use 720 sentences as training data and the rest of the recordings as test data in our speaker dependent experimental evaluations. The recordings are phonetically transcribed using Turkish phonetic dictionary METUbet [15], and the phone level alignment is performed using forced-alignment.

In our recent study [13], we report our findings for the benefit of using phone-dependent probabilistic mappings to enhance TM spectral envelope. In this study we consider only phone-dependent mapping schemes with two possible sources of the phone. First source is obtained by force alignment, and the second source is obtained from an HMM-based phone recognition system over the observable TM source. The phone recognition performances for the M1 and M2 speakers are obtained respectively as 62.22% and 61.07%. In the phone-dependent probabilistic mapping we use \( L = 16 \) Gaussian mixture components for each phone class, since we do not observe a significant performance difference for a range of mixtures from 8 to 32.

3.1. Objective Evaluations

The ITU-T Standard PESQ [16] is employed as the objective metric to evaluate the perceptual quality of the enhancement of TM recordings. We first investigate possible best case scenarios for the enhancement of TM recordings when AM recordings are available. Table 1 presents average PESQ scores for the four scenarios, where reference condition is always the AM recordings and a form of the TM recordings are synthesized with the given excitation and filter models. The first row presents the average PESQ scores between TM and AM recordings for both speakers M1 and M2. The second row presents the average PESQ scores of the source-filter synthesis when the TM filter is replaced by the AM filter. We observe similar PESQ score improvements for both speakers, and these improvements can be considered as best case improvements for the enhancement of the spectral envelope. The last two rows of Table 1 presents average PESQ scores when AM filter is used and TM excitation is tilted with the original spectral tilt vector with linearly spaced frequency bands (\( D_{lin} \)) and with mel-scaled frequency bands (\( D_{mel} \)). Note that when the TM excitation is tilted with the original spectral tilt vector we observe high PESQ score improvements. However we do not observe any improvement with the use of mel-scaled frequency bands in the computation of the spectral tilt vectors. Hence we keep using the linearly spaced frequency bands in the remaining parts of the experimental evaluations.

Table 1: The average PESQ scores for evaluation of the targeted excitation and filter enhancement strategies.

<table>
<thead>
<tr>
<th></th>
<th>Exc</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_T )</td>
<td>1.22</td>
<td>1.42</td>
</tr>
<tr>
<td>( R_T )</td>
<td>1.70</td>
<td>1.74</td>
</tr>
<tr>
<td>( R_A^{lin} )</td>
<td>2.23</td>
<td>2.72</td>
</tr>
<tr>
<td>( R_A^{mel} )</td>
<td>2.26</td>
<td>2.63</td>
</tr>
</tbody>
</table>

Figure 1: Sample spectrograms of (a) AM excitation, (b) enhanced TM excitation with the original spectral tilt, (c) enhanced TM excitation with the estimated spectral tilt, and (d) TM excitation.
Figure 1 presents spectrograms of the excitation signals to emphasize the effect of spectral tilt. The top plot is the AM excitation and the bottom plot is the TM excitation spectograms. The second from the top (b) is the TM excitation, which is enhanced by the original spectral tilt vector. Similarly, the third from the top (c) is the TM excitation, which is enhanced by the estimated spectral tilt vector using the proposed probabilistic mapping. Note that enhancement of the TM excitation significantly compensates spectral energy distribution with respect to the AM excitation.

Table 2 presents PESQ scores for different excitation and spectral envelope mapping schemes that we use for the enhancement of TM recordings. First two columns define excitation and filter mapping schemes. The PESQ scores are presented with respect to the source of phone information, either force alignment or HMM-based phone recognition, and speaker information M1 and M2. The first observation is on the source of the phone information. Given the PESQ scores with the reliable force aligned phone information, the PESQ scores with the HMM-based phone information do not degrade significantly. For example, the first row of the results, where only the spectral envelope is mapped, PESQ score drop is from 1.59 to 1.58 for speaker M2 with the use of phone recognition information. The second observation is on the attained PESQ score improvements with the proposed excitation and spectral envelope mapping schemes. The PESQ scores 1.22 and 1.42 of the TM recordings as reported in Table 1 respectively for speakers M1 and M2 are improved to 1.46 and 1.58 with the spectral envelope mapping, and to 1.61 and 1.86 with the excitation and spectral envelope mappings when phone information is extracted from the HMM-based phone recognizer. Here surprising observation is the contribution of the excitation enhancement, which brings a higher improvement than the spectral envelope enhancement. Finally the third observation is on the sole contribution of the excitation mapping, where last row of Table 2 presents PESQ scores with the proposed excitation mapping when TM filter is used. We do not observe PESQ score improvements for the sole use of excitation mapping, hence we can say that the proposed excitation mapping is significant when used with the spectral envelope mapping.

<table>
<thead>
<tr>
<th>Exc</th>
<th>Filter</th>
<th>Force Alignment</th>
<th>Phone Recog</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_T$</td>
<td>$W_A$</td>
<td>1.53</td>
<td>1.59</td>
</tr>
<tr>
<td>$\hat{R}_T$</td>
<td>$W_A$</td>
<td>1.65</td>
<td>1.90</td>
</tr>
<tr>
<td>$\hat{R}_T$</td>
<td>$W_T$</td>
<td>1.34</td>
<td>1.43</td>
</tr>
</tbody>
</table>

Table 2: The average PESQ scores for different excitation and spectral envelope mapping schemes.

3.2. Subjective Evaluations

We performed a subjective A/B comparison test to evaluate the proposed enhancement techniques. During the test, the subjects are asked to indicate their preference for each given A/B test pair of sentences on a scale of (-2; -1; 0; 1; 2), where the scale corresponds to strongly prefer A, prefer A, no preference, prefer B, and strongly prefer B, respectively. The subjective A/B test includes 33 listeners, who compared 29 sentence pairs randomly chosen from our test database of speaker M2 to evaluate 6 conditions. The AM and TM speech conditions are compared to all conditions with 1 pair. The second row of the Table 1 ($R_T$, $\hat{W}_A$) defines a target condition for the best possible spectral envelope mapping, which is evaluated as the third condition. The fourth condition is defined as the target for the best possible excitation and spectral envelope mapping ($\hat{R}_T$, $\hat{W}_A$), which is the third row of the Table 1. The fifth and sixth conditions are set as the two proposed enhancement schemes, the sole spectral envelope mapping ($\hat{R}_T$, $\hat{W}_A$) and the excitation and spectral envelope mappings ($\hat{R}_T$, $\hat{W}_A$), which are respectively the first two rows of the Table 2. Conditions 3-6 are compared to each other with 3 pairs.

Table 3 presents the average subjective preference results. The rows and the columns of Table 3 correspond to A and B conditions of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection. Speech samples from the subjective A/B comparison test are available for online demonstration [17].

The proposed excitation and spectral envelope mapping scheme ($\hat{R}_T$, $\hat{W}_A$), which is the condition 6, has perceivable improvements compared to all conditions except the AM recordings and the best target mapping ($\hat{R}_T$, $\hat{W}_A$). Furthermore it is significantly preferred over the sole spectral envelope mapping ($\hat{R}_T$, $\hat{W}_A$) with a preference score 1.12, and it is the second closest condition to the AM recordings after the best target mapping with a preference score -1.49.

Table 3: The average preference results of the subjective A/B pair comparison test.

4. Conclusions

We introduce a new phone-dependent probabilistic mapping scheme for the excitation signal to enhance TM speech using joint analysis of TM and AM recordings. The proposed excitation mapping scheme performs the MMSE estimation of the spectral difference between AM and TM excitation spectrums within the phone class neighborhoods. Objective and subjective experimental evaluations indicate that the correction of the TM excitation spectrum has a strong potential to improve intelligibility for the TM speech. Furthermore the proposed phone-dependent excitation mapping yields perceivable improvements over the state-of-the-art TM speech enhancement schemes. Although the proposed excitation mapping achieves a significant improvement (1.86 PESQ score for speaker M2) within the strong potential of correcting TM excitation with a spectral tilt (2.72 PESQ score of the best target mapping for speaker M2), there is still a room for further improvement. We consider that incorporating temporal dynamics of the spectral tilt to the probabilistic mapping may attain further improvements for TM speech enhancement.
5. References


