Quality Conversion of Non-Acoustic Signals for Facilitating Human-to-Human Speech Communication under Harsh Acoustic Conditions

Seyed Omid Sadjadi, Sanjay A. Patil, and John H.L. Hansen

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA
{sadjadi, sanjay.patil, john.hansen}@utdallas.edu

Abstract

Harsh acoustic conditions limit the effectiveness of human speech communication to a great extent. There is a consensus that even at moderate SNR levels, traditional speech enhancement techniques tend to improve the perceptual quality of speech rather than its intelligibility. As an alternative, non-acoustic contact sensors have recently been developed for noise-robust signal capture. Although relatively immune to ambient noise, due to alternative pickup location and non-acoustic principle of operation, signals measured from these sensors are of lower speech quality and intelligibility when compared to those obtained from a conventional microphone in clean conditions. To facilitate human-to-human speech communication under acoustically adverse environments, in this study we present and evaluate a probabilistic transformation framework to improve perceptual quality and intelligibility of signals acquired from one such sensor entitled: physiological microphone (PMIC). Results from both objective and subjective tests confirm that incorporating this framework as a post-processing stage yields significant improvement in overall quality and intelligibility of the PMIC signals.

Index Terms: physiological microphone, probabilistic transformation, quality conversion, speech perception, speech quality

1. Introduction

Speech is the primary means of communication among humans. There are a number of scenarios where effective human-to-human speech communication is vital, yet adverse noisy conditions limit intelligible information exchange and degrade quality. The degraded speech quality reduces acoustic cue redundancy [1] and causes an increase in the cognitive load of the listener, often referred to as listener fatigue. Traditionally, front-end speech enhancement techniques have been employed to alleviate the impact of ambient noise on both human-human and human-machine speech communication. However, it is well known that because these enhancement techniques have been designed to primarily improve speech quality by suppressing background noise, they provide limited or generally no improvement in intelligibility [2].

In addition, while such approaches appear to improve the degraded signal quality at moderate signal-to-noise ratios (SNR), they may introduce annoying distortions (e.g., artifacts and musical noise) for extremely low SNR scenarios where the signal is completely obscured in noise (e.g., speech communication among firefighters or workers in a factory environment with heavy-metal manufacturing facilities).

Consequently, as an alternative to the conventional microphone technology, the development and utilization of non-acoustic contact sensors for noise-robust speech systems have attracted significant research effort in the past decade [3]–[7]. This activity is motivated because these sensors are essentially independent of acoustic characteristics of environments under which they are operating (e.g., background noise and reverberation).

Although robust against acoustic disturbances, due to alternative pickup location and non-acoustic principles of operation, signals measured from these sensors do not reflect the same lip radiation and vocal tract resonance properties, and their spectral content suffer from frequency characteristics of the body tissue. Furthermore, because these sensors record the glottal activity that occurs during phonation, the unvoiced portion of speech is only weakly represented in their signals. These issues give rise to reduced speech quality and intelligibility of the signal acquired from non-acoustic sensors, when compared to that acquired from traditional microphones (e.g., close-talk microphone (CTM)). As a consequence an additional post-processing stage would be useful to enhance quality and intelligibility of their signals before delivering them to listeners, which could be either machines or humans.

Specifically, in this study we focus on the use of one such sensor entitled: “physiological microphone (PMIC)” [3], and present a continuous probabilistic transformation approach as the post-processing stage. Our objective is to improve overall quality and intelligibility of signals acquired from the PMIC for human listeners, in order to reduce listening effort and thus facilitate a more robust human-to-human speech communication experience under harsh acoustic conditions. Our approach involves mapping the acoustic space of the PMIC into conventional speech production space represented by signals measured from a CTM. Both the PMIC and CTM acoustic spaces are modeled by codebooks from vector quantization (VQ) of the spectral envelopes obtained from a small amount of parallel training data. The transformation is performed piecewise linearly using a continuous parametric function which takes into account probabilistic classification of the acoustic spaces provided by a Gaussian mixture model fitted on the codebooks. Parameters of the transformation function are determined by solving a linear least squares estimation problem for each VQ cluster. Both the PMIC and CTM acoustic spaces are modeled by codebooks from vector quantization (VQ) of the spectral envelopes obtained from a small amount of parallel training data. The transformation is performed piecewise linearly using a continuous parametric function which takes into account probabilistic classification of the acoustic spaces provided by a Gaussian mixture model fitted on the codebooks. Parameters of the transformation function are determined by solving a linear least squares estimation problem for each VQ cluster. Both the PMIC and CTM acoustic spaces are modeled by codebooks from vector quantization (VQ) of the spectral envelopes obtained from a small amount of parallel training data. The transformation is performed piecewise linearly using a continuous parametric function which takes into account probabilistic classification of the acoustic spaces provided by a Gaussian mixture model fitted on the codebooks. Parameters of the transformation function are determined by solving a linear least squares estimation problem for each VQ cluster. Both the PMIC and CTM acoustic spaces are modeled by codebooks from vector quantization (VQ) of the spectral envelopes obtained from a small amount of parallel training data. The transformation is performed piecewise linearly using a continuous parametric function which takes into account probabilistic classification of the acoustic spaces provided by a Gaussian mixture model fitted on the codebooks. Parameters of the transformation function are determined by solving a linear least squares estimation problem for each VQ cluster.

2. Acoustic properties of PMIC signals

The PMIC is a non-acoustic contact sensor which captures a surface vibration from movement of speech production organs during phonation [3]. When strapped around the throat near the carotid and thyroid cartilages (as shown in Fig. 1), it gives the best performance and provides good noise attenuation (~ 30 dB) as well as good excitation information [3], [4]. Due to the impedance matching property
of the sensor-body and its low air-borne acoustic coupling index, the PMIC is virtually immune to acoustic noise.
To provide a clear picture of acoustic properties of PMIC signals, the long-term average speech spectrum (LTASS) [8] obtained for 10 sentences (about 30 seconds) simultaneously recorded from an adult male speaker using the PMIC and CTM is depicted in Fig. 2. There are some important features in the PMIC LTASS when compared to that of the CTM:
- It shows appreciable energy below 0.5 kHz, indicating that there is strong voicing information available in PMIC signals since it is located close to the vocal folds where it can pick up source excitation faithfully.
- It does not show the same typical spectral tilt seen in the CTM spectrum, pointing to the fact that lip radiation effect is not reflected in PMIC signals.
- It has a fast roll-off for frequencies above 2.7 kHz. This is due to the fact that: (i) the body tissue acts as a lowpass filter and attenuates spectral energy in higher frequency regions, (ii) the PMIC basically captures the glottal activity that occurs during phonation, therefore the unvoiced portion of speech is weakly presented in its signals.
- Although the first and second formants appear to be preserved in the PMIC spectrum, their locations and bandwidths as well as their magnitudes have been changed. Generally speaking, for frequencies above 0.6 kHz the PMIC spectrum exhibits weaker energy compared to the CTM spectrum.

![Figure 1: The PMIC (right) and its position around the throat (left).](image)

Many studies have shown the impact of changes in typical shape of speech spectrum on intelligibility and quality. In particular, listening experiments have revealed that the acoustic spectral information, e.g., formant locations, formant bandwidth, spectral tilt, and spectral contrast, play a significant role in perceived quality and intelligibility of speech signals [8], [9], and [1]. This along with the above observations point to a decline in perceptual quality and intelligibility of PMIC signals, which reflects our prime motivation in formulating an approach to improve these aspects. The approach should be able to minimize the mismatch between the acoustic spaces of signals measured by each sensor technology to achieve comparable signal qualities. Such an approach is proposed in the next section.

### 3. Probabilistic transformation algorithm

In this section, we first formulate the probabilistic transformation algorithm, and then consider implementation for mapping the PMIC signal into the conventional CTM speech response.

#### 3.1. Problem formulation

Assuming that parallel signal data are available from the PMIC and CTM, our goal is to minimize the mismatch between the spectrum response of the two sensors represented by spectral envelopes extracted from corresponding recordings. Our expectation is that in this manner PMIC signal quality will move to that perceived from a comparable CTM signal. More precisely, let \( \{x_t, t = 1, \ldots, N\} \) and \( \{y_t, t = 1, \ldots, N\} \) denote two sets of corresponding PMIC and CTM \( n \)-dimensional vectors of line spectral frequencies (LSF), representing the spectral envelopes (\( N \) is the total number of training frames). A linear mapping function \( \Phi(\cdot) \) that transforms each vector of the set \( \{x_t\} \) into its counterpart in the set \( \{y_t\} \) is defined as,

\[
\Phi(x_t) = A^T x_t + b, \quad (1)
\]

where \( A \) is an \( n \times n \) matrix, and \( b \) is an \( n \times 1 \) additive term, both of which are to be estimated. Rearranging (1) into a single matrix equation, we obtain

\[
\Phi(x_t) = H \tilde{x}_t, \quad (2)
\]

where,

\[
\tilde{x}_t = [x_t \quad 1]^T, \quad H = [A^T \quad b].
\]

Now, \( \Phi(\cdot) \) is entirely defined by the \( (n+1) \times n \) transformation matrix \( H \). This matrix is computed by least squares optimization on parallel training data to minimize the sum of squared transformation errors,

\[
\epsilon = \sum_{t=1}^{N} \|y_t - \Phi(x_t)\|^2. \quad (3)
\]

The optimal solution for the matrix \( H \) is given by [10],

\[
H_{opt} = R_{yx} R_{xx}^{-1}, \quad (4)
\]

in which \( R_{xx} = \sum_{t=1}^{N} \tilde{x}_t \tilde{x}_t^T \) is the autocorrelation matrix of vectors in the set \( \{x_t\} \), and \( R_{yx} = \sum_{t=1}^{N} y_t \tilde{x}_t^T \) is the crosscorrelation matrix of vectors in the two sets.

To gain a better interpolation, thus compensating for possible losses in acoustic information discussed in Section 2, the feature vector \( \tilde{x}_t \) can be modified to \( \tilde{X}_t = [x_{t-1}^T \ldots x_t^T \ldots x_{t+k}^T]_T \), with \( k \) being the number of frames in the neighborhood of the current frame \( t \). Moreover, in order for \( \Phi(\cdot) \) to perform well across different phoneme classes within speech, it should take into account the classification of the PMIC acoustic space. This way, the general transformation matrix \( H \), turns into a class specific one, and the transformation can then be performed piecewise linearly.

Classification can be realized through a VQ framework, however the discrete nature (i.e., one-to-one mapping) inherent in VQ may hurt the transformation quality. As a remedy, we fit a GMM on the performance of the two sensors represented by spectral envelopes extracted from corresponding recordings. Our expectation is that in this manner PMIC signal quality will move to that perceived from a comparable CTM signal. More precisely, let \( \{x_t, t = 1, \ldots, N\} \) and \( \{y_t, t = 1, \ldots, N\} \) denote two sets of corresponding PMIC and CTM \( n \)-dimensional vectors of line spectral frequencies (LSF), representing the spectral envelopes (\( N \) is the total number of training frames). A linear mapping function \( \Phi(\cdot) \) that transforms each vector of the set \( \{x_t\} \) into its counterpart in the set \( \{y_t\} \) is defined as,

\[
\Phi(x_t) = A^T x_t + b, \quad (1)
\]

where \( A \) is an \( n \times n \) matrix, and \( b \) is an \( n \times 1 \) additive term, both of which are to be estimated. Rearranging (1) into a single matrix equation, we obtain

\[
\Phi(x_t) = H \tilde{x}_t, \quad (2)
\]

where,

\[
\tilde{x}_t = [x_t \quad 1]^T, \quad H = [A^T \quad b].
\]

Now, \( \Phi(\cdot) \) is entirely defined by the \( (n+1) \times n \) transformation matrix \( H \). This matrix is computed by least squares optimization on parallel training data to minimize the sum of squared transformation errors,

\[
\epsilon = \sum_{t=1}^{N} \|y_t - \Phi(x_t)\|^2. \quad (3)
\]

The optimal solution for the matrix \( H \) is given by [10],

\[
H_{opt} = R_{yx} R_{xx}^{-1}, \quad (4)
\]

in which \( R_{xx} = \sum_{t=1}^{N} \tilde{x}_t \tilde{x}_t^T \) is the autocorrelation matrix of vectors in the set \( \{x_t\} \), and \( R_{yx} = \sum_{t=1}^{N} y_t \tilde{x}_t^T \) is the crosscorrelation matrix of vectors in the two sets.

To gain a better interpolation, thus compensating for possible losses in acoustic information discussed in Section 2, the feature vector \( \tilde{x}_t \) can be modified to \( \tilde{X}_t = [x_{t-1}^T \ldots x_t^T \ldots x_{t+k}^T]_T \), with \( k \) being the number of frames in the neighborhood of the current frame \( t \). Moreover, in order for \( \Phi(\cdot) \) to perform well across different phoneme classes within speech, it should take into account the classification of the PMIC acoustic space. This way, the general transformation matrix \( H \), turns into a class specific one, and the transformation can then be performed piecewise linearly.

Classification can be realized through a VQ framework, however the discrete nature (i.e., one-to-one mapping) inherent in VQ may hurt the transformation quality. As a remedy, we fit a GMM on the codeword obtained from the VQ and incorporate the probabilistic classification (soft decision making) of the entries within the transformation by updating (3) as,

\[
\epsilon_{q} = \sum_{t=1}^{N} \|y_t - H_{q}^T \tilde{x}_t\|^2, \quad (5)
\]

where \( p(C_q|x_t) \) is the posterior probability of the \( q \)-th class \( C_q \), given the input vector \( x_t \), and \( H_q \) is the \((2k+1)(n+1) \times n \) transformation matrix for \( C_q \) which is
computed as in (4), with $R_{xx} = \sum_{t=k}^{N-k} \tilde{X}_t \tilde{X}_t^T p(C_q|x_t)$ and $r_{xy} = \sum_{t=k}^{N-k} \tilde{y}_t \tilde{X}_t^T p(C_q|x_t)$. The transformed feature vector is then calculated by summing the outputs of all the classes as [11],

$$\hat{y}_t = \Phi(x_t) = \sum_{q=0}^{Q-1} p(C_q|x_t) H_q^T \tilde{X}_t,$$

where $Q$ denotes the total codebook size.

Based on the above formulation, two alternative transformation strategies can be distinguished. The first strategy is the well-known codebook mapping (plain VQ), where the index of the PMIC codebook entry with the minimum distance from the input vector $x_t$ is used to extract the corresponding CTM entry, $Y_q$. Now, if we weight the CTM codebook entries, $Y_q$, with the posterior probability of the PMIC vectors, $p(C_q|x_t)$, and sum across all the clusters, the second transformation method is obtained as,

$$\hat{y}_t = \sum_{q=0}^{Q-1} p(C_q|x_t) Y_q,$$

We will refer to this second method as probabilistic VQ.

3.2. Implementation

In the training phase, after the LSFs are extracted from the available parallel training data, the CTM space is vector quantized into $Q$ clusters using the well-known binary split LBG algorithm. Consequently, a GMM is fitted on the corresponding PMIC codebook, and the posterior probabilities $p(C_q|x_t)$ are calculated. The length of the segment feature $\tilde{X}_t$ is fixed at 5 feature frames ($k = 2$), and the transformation matrix $H_q$ is computed. During the transformation phase, the enhanced PMIC spectral envelopes are calculated via (6), and convolved with the corresponding excitation signal extracted from the original PMIC recording. It is worth noting that in the present system, the problem of transforming the source excitation is not considered since it is reliably captured by the PMIC (also see Section 2 and [4]).

Figure 3: The long-term average speech spectrum (LTASS) obtained from about 30 seconds of the CTM, PMIC, and transformed PMIC signals.

4. Results and discussion

The proposed approach is evaluated on 60 sentences (6 sentence lists) from the IEEE database [12] recorded simultaneously from an American adult male speaker using the PMIC and the CTM. The IEEE sentences are phonetically balanced with relatively low word-context predictability, which makes it well suited for subjective quality and intelligibility tests. In a leave-one-out (LOO) cross validation scheme, from a total of 6 sentence lists, sentences from 5 lists are used to train the transformation matrix, while the remaining sentence set is employed for evaluation. The LOO strategy is adopted to ensure that the results are not biased due to data partitioning.

4.1. Objective tests

Fig. 3 displays the LTASS obtained from the same data used to generate Fig. 2, for the PMIC, CTM, and transformed signals. The log-spectral distortion measure between the original PMIC and the CTM spectra is 15.99 dB, while it is 0.51 dB between the transformed PMIC and the CTM. It is also evident from the figure that the proposed approach has reliably minimized the mismatch between the two spectral responses.

In order to illustrate performance of the proposed algorithm in enhancing the spectral information of PMIC signals with more details for different phoneme classes, spectrograms and time waveforms obtained from a sample sentence are shown in Fig. 4. It is clear that the PMIC spectrogram is dominated by low frequency components (below 0.6 kHz), however, a weak activity replicating the CTM spectrogram can be observed for higher frequencies up to 3 kHz after which there is almost no visible spectral component. In addition, the temporal structure of the speech signal is altered to a great extent for the PMIC when compared to the CTM, and the change is more dramatic for consonants than vowels. Comparing the three signal types in time and frequency, it can be seen that the transformation system is able to effectively reconstruct the attenuated spectral contents as well as the deformed temporal structure, for different phoneme classes.

Fig. 5 compares the performance of the proposed probabilistic transformation (PT) approach against that of the other two transformation strategies introduced in Section 3.1, in terms of the perceptual evaluation of speech quality (PESQ) and log-likelihood ratio (LLR) quality metrics. Both of these metrics have been shown to be highly correlated with the subjective listening tests for speech quality [13]. The bars labeled as “UN” show the metrics obtained from comparing the original PMIC (unprocessed) and CTM signals. It is obvious from the figure that the proposed approach outperforms the other two strategies, albeit with more computational load. Moreover, increasing the codebook size does help the performance of all the methods, however it does not apply to the proposed approach when the size goes beyond 7 bits, where a drop in performance is observed. This happens due to data scarcity in some VQ clusters which makes the autocorrelation matrix $R_{xx}$ ill-conditioned, and
consequently the resulting estimates will be unstable. Furthermore, introducing the soft decision making into the plain VQ method (VQ-Plain) results in a measurable boost seen in the performance shown for the probabilistic VQ (VQProb) technique.

4.2. Subjective listening tests
A total of 14 normal-hearing American listeners were recruited, paid, and trained for the listening experiments. The experiment was performed in a sound proof room. Two lists of sentences (20 sentences) were used per signal type (i.e., the original PMIC, processed PMIC, and CTM). Stimuli were played to the listeners through open-air Sennheiser HD580 headphone at their desired volume levels. Subjects were asked to judge overall speech quality as well as their own listening effort. They were also asked to transcribe what they heard. The appearance of sentences from different signal types was randomized to ensure that the results are not biased due to listeners predictions.

Fig. 6 demonstrates the results of the listening experiment in terms of perceived quality and listening effort (left half) as well as word error rate (WER) scores (right half), averaged across the subjects. As can be seen from the left half of the figure (primary vertical axis), there is a dramatic elevation in overall quality of the enhanced PMIC signals compared to the original recordings. The amount of listening effort is also relatively decreased for the processed signals. As expected, WER scores (secondary vertical axis) indicate a decline in intelligibility of the PMIC recordings compared to that of the CTM speech. It is observed that the proposed approach has successfully improved the intelligibility of the original PMIC signals. In all the cases, improvements are statistically significant (paired t-test: $p < 0.001$).

5. Conclusions

The present study proposed a continuous probabilistic transformation technique for improving overall quality and intelligibility of signals measured from the non-acoustic PMIC. The technique was objectively benchmarked against two conventional transformation strategies, and shown to be superior in performance. Subjective listening experiments were carried out to assess (i) overall quality, (ii) listening effort, and (iii) intelligibility of the transformed PMIC signals, and the outcomes indicated the probabilistic transformation method is quite effective in removing the metallic and muffled nature prevalent in PMIC signals, as well as increasing intelligibility.

6. References