TOWARDS MORE INTELLIGIBLE PHYSIOLOGICAL MICROPHONE SPEECH: A PROBABILISTIC TRANSFORMATION APPROACH

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

The Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX 75080, USA
sadjadi@student.utdallas.edu, {sanjay.patil, john.hansen}@utdallas.edu

ABSTRACT
The non-acoustic physiological microphone (PMIC) has been shown to be useful for speech systems under adverse noisy conditions. However, the signal is not a true speech for the listener, therefore appears muffled and metallic with variations to the speaker dependent structure. This study presents a probabilistic transformation approach to improve the perceptual quality and intelligibility of PMIC speech not only by mapping the non-acoustic signal into the conventional speech production space, but also by minimizing distortions arising from alternative pickup location. Performance of the proposed approach is assessed based on five distinct objective metrics. Obtained results indicate that incorporating the probabilistic transformation yields significant improvement in overall PMIC speech quality and intelligibility. This technique along with the PMIC can thus find applications in noise robust human-to-human speech communication.

Index Terms— Linear mapping, physiological microphone, probabilistic transformation, speech quality, objective quality measure

1. INTRODUCTION
Adverse noisy conditions can pose problems for both human-human and human-machine speech communication, since they lead to reduced signal quality and intelligibility [1]. For example, in speech communication among workers in a factory environment with heavy metal manufacturing facilities, overall signal quality is severely affected by the presence of highly non-stationary noise resulting in increased listening effort. This in turn demands higher cognitive load on the listener which diminishes human performance. Traditionally, front-end speech enhancement techniques have been employed to alleviate the impact of noise on the performance of both man and machine. While such techniques appear to improve the degraded signal quality, they may introduce artifacts, and for extremely low SNR scenarios provide limited or generally no improvement.

An alternative approach to deal with these scenarios for making speech communication more robust against environment noise is the use of non-acoustic sensors, which are essentially independent of acoustic background noise. Bone-conductive microphone has been successfully applied to multi-sensory speech enhancement [2]. Glottal Electromagnetic Micropower Sensor (GEMS) and the PMIC have been exploited in multi-sensor low-rate coders as well as in background noise removal frameworks for improved speech quality and intelligibility under heavy noisy conditions [3], [4]. The PMIC has also been shown to be quite effective for ASR tasks, especially at relatively low SNR situations where the use of conventional microphone technology fails [5]. In a more recent attempt for noise-robust speech communication, Non-Audible Murmur (NAM) microphone along with statistical voice conversion approaches for enhancing the body transmitted speech have been employed [6].

In this paper, we focus on the use of the PMIC sensor for noise-robust signal capture. Although the PMIC signal is robust against environment noise, it is not a true speech for the listener, therefore appears muffled and metallic. In other words, it is less intelligible than conventional close-talk microphone (CTM) speech. This is due to its non-acoustic principle of operation and bandstop like characteristics of the PMIC-skin interface. As a consequence, along with the PMIC, an additional post processing stage would be useful for enhancing overall speech quality and intelligibility by reconstructing those spectral components which have suffered maximum loss due to alternative signal pickup location.

This study proposes a probabilistic transformation approach as the post processing stage. The approach was originally exploited in [7] to map the noisy speech feature space into the clean speech feature space, for robust speech recognition. Assuming the linear source-filter speech production model, our approach involves mapping the filter parameters, (i.e., the spectral envelope and gain) of the PMIC signal to estimate the CTM speech. The mapping is performed piecewise linearly based on a pre-trained vector quantization (VQ) codebook using only a small amount of parallel training data. However, instead of performing a hard-decision (i.e., one-to-one mapping) as implemented in a traditional VQ, we incorporate a probabilistic soft-decision strategy within the VQ framework, which helps reduce artifacts associated with the mapping process.

The paper is organized as follows: In Section 2, a brief description of the PMIC sensor is provided and the speech signal acquired in this scenario is compared against that obtained from the CTM. The next section gives details on the proposed probabilistic transformation approach, followed by details on the experiments conducted in Section 4. Section 5 presents results and analysis. Finally, we draw conclusions and outline potential applications based on our approach in Section 6.

Fig. 1. The PMIC sensor (right) and its position around the throat (left).

2. PMIC VS. CTM
The PMIC is a non-acoustic contact sensor capable of capturing speech related signals through skin vibrations pickup [5]. When strapped around the throat near the carotid and thyroid cartilages (as

⋆This project was supported in part by USAF under a subcontract to RADC, Inc. under FA8750-05-C-0029. Approved for public release; distribution unlimited.
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. In this study, we focus on the "i am" phrase (1100-1400 ms), the CTM spectrum shows strong consistent in the PMIC signal spectrum. At the end of the utterance, for 600-1400 ms interval. The frequency band 2-3 kHz is almost nonexistent in the PMIC signal spectrum. As depicted in Fig. 2, the PMIC signal possesses predominantly lower frequency (less than 1 kHz) or higher frequency (greater than 3 kHz) content, however weak activity replicating the CTM spectrogram can also be observed between 1 kHz and 3 kHz. The spectral content for the fricative /sh/ (100-200 ms interval) is clearly visible from the CTM spectrogram but appears very weak for the PMIC. Moreover, formant location and bandwidth changes occur within 600-1400 ms interval. The frequency band 2-3 kHz is almost nonexistent in the PMIC signal spectrum. At the end of the utterance, for the "i am" phrase (1100-1400 ms), the CTM spectrum shows strong formant transition while the PMIC is unable to track this structure.

Listening experiments have revealed that the acoustic spectral information (e.g., formant locations, formant bandwidth, and spectral tilt) play a significant role in perceived speech quality [1], [8]. This along with the above observations point to the lack of speech quality and intelligibility in the PMIC signal, which reflects our prime motivation to formulate an approach to improve these aspects. The approach should be able to reconstruct, and if possible strengthen, weak formant peaks, thus improving overall vocalic qualities [1]. 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. Mathematical formulation

A general function \( \psi \) for estimating a source feature vector \( \mathbf{x}_t = [x_0 \ldots x_{n-1}]^T \) from a target feature vector \( \mathbf{y}_t = [y_0 \ldots y_{m-1}]^T \) at frame \( t \) is defined as,

\[
\mathbf{y}_t = \psi(\mathbf{x}_t). \tag{1}
\]

For the linear estimation case, (1) becomes,

\[
\hat{\mathbf{y}}_t = \mathbf{H}^T \mathbf{x}_t. \tag{2}
\]

where \( \mathbf{H} \) can be interpreted as an FIR filter-coefficient matrix, which maps the source feature vector into the target feature space. This matrix is computed by minimizing the sum of the squared estimation errors \( \varepsilon = \sum |e_t|^2 \), where the error is defined as the difference between the target and the estimated vectors,

\[
e_t = y_t - \hat{y}_t = y_t - \mathbf{H}^T \mathbf{x}_t. \tag{3}
\]

The optimal solution [9] for the transformation matrix \( \mathbf{H} \) is given by

\[
\mathbf{H}_{opt} = \mathbf{R}^{-1} \mathbf{r}, \tag{4}
\]

in which \( \mathbf{R} = \mathbf{x}_t \mathbf{x}_t^T \) is the autocorrelation matrix of the target vector and \( \mathbf{r} = \mathbf{y}_t \mathbf{x}_t \) is the crosscorrelation matrix of the source and target vectors.

For \( \mathbf{x}_t \) and \( \mathbf{y}_t \) as the given speech feature vectors, a single universal solution via (4) will not perform well across all speech sound groups. This is primarily due to the diversity in the speech phoneme inventory, demanding a pre-clustering stage which can be realized through VQ processing phase (i.e., performing the mapping piecewise linearly). Moreover, because of the variations in a phoneme arising from coarticulation, the phoneme will have a certain probability of belonging to a VQ cluster. Incorporating this probability into the error minimizing function \( \varepsilon \) for each cluster, we can obtain,

\[
\varepsilon_q = \sum |e_{tq}|^2 p(C_q|z_t), \tag{5}
\]

where \( p(C_q|z_t) \) is the posterior probability of the \( q \)-th cluster, \( C_q \), given the conditioning source feature vector \( z_t \), and \( e_{tq} \) is the estimation error associated with that cluster.

Furthermore, due to the slowly varying nature of speech signals, feature vectors across frames will in general be inter-dependent. Therefore, incorporating this inter-frame dependency into the transformation framework will yield better interpolation and feature smoothing. Taking all of the above facts into account, the source feature vector is redefined as \( \mathbf{X}_t = [x_{t-k} \ldots x_t \ldots x_{t+k}]^T \), with \( k \) being the number of frames in the neighborhood of the current frame, and (3) and (4) are rewritten as,

\[
e_{tq} = y_t - \hat{y}_t = y_t - \mathbf{H}^T \mathbf{X}_t, \tag{6}
\]

\[
\mathbf{H}_{q,opt} = \mathbf{R}_{q}^{-1} \mathbf{r}_{q}, \tag{7}
\]

where \( \mathbf{R}_{q} = \sum_{t=k}^{N-k-1} \mathbf{X}_t \mathbf{X}_t^T \) is a probabilistic autocorrelation matrix while \( \mathbf{r}_{q} = \sum_{t=k}^{N-k-1} \mathbf{y}_t \mathbf{x}_t^T \) is a probabilistic crosscorrelation matrix with \( N \) being the total number of training frames. The estimate of the target feature vector \( \hat{y}_t \), is then calculated by summing the outputs of all the VQ clusters as,

\[
\hat{y}_t = \sum_{q=0}^{Q-1} \mathbf{H}_{q,opt} \mathbf{X}_t p(C_q | z_t), \tag{8}
\]

where \( Q \) denotes the codebook size.

3.2. Implementation

Linear source-filter theory of speech production models speech in terms of three independent parameters: an excitation source, a vocal tract filter, and a gain factor. Exploiting the ability of the throat-strapped PMIC to capture the excitation information faithfully, in this study we use the vocal tract filter parameters and gain factor are considered in order to train the transformation while the source excitation from the PMIC signal is used to resynthesize the CTM speech signal. Here, line spectral frequencies (LSFs) obtained from a 12th-order LPC analysis are used to represent the filter. LSFs are stable and have good linear interpolation properties. In addition, they are closely related to formant locations and bandwidths. These qualities make LSFs the ideal feature for the current task involving interpolation and enhancement of the degraded frequency components of PMIC signal.

In the training phase, after the LSFs and gain feature set is extracted from parallel training database of PMIC and CTM speech signals, the CTM feature space is vector quantized into \( Q \) clusters using the binary split LBG algorithm [10]. Subsequently, the posterior probability of each cluster, \( p(C_q | z_t) \), is calculated using Bayes’
theorem as follows,
\[ p(C_q | z_t) = \frac{p(C_q) p(z_t | C_q)}{\sum_{q=0}^{Q} p(C_q) p(z_t | C_q)} \]

in which \( p(z_t | C_q) \) is modeled as a mixture of \( Q \) Gaussian distributions with diagonal covariance matrices. \( p(C_q) \) is the prior probability of cluster \( C_q \) estimated based on relative frequency from the training data. The conditioning feature vector \( z_t \) can be any acoustic parameter extracted from the PMIC data (e.g., LSF, MFCC, LPCC). For this study, in order to avoid the computational expense of introducing a new set of speech features, 12-dimensional LSFs are also employed as \( z_t \). The length of \( Z_t \) is fixed at 5 feature frames (\( k = 2 \)). Finally, two independent transformation matrices, \( H_{1,\text{opt}} \), are computed for mapping the LSFs and the gain factor. The gain estimation is crucial in mapping PMIC to CTM as it plays an important role in shaping the overall synthesized speech waveform. Reliable estimation of this parameter can amplify those parts of the PMIC signal attenuated due to the distortion caused by the alternative pickup location.

During the estimation phase, the estimated CTM feature vectors are calculated via (8), and the transformed speech signal is resynthesized by exciting the estimated vocal tract filter with the corresponding source signal extracted from the input PMIC signal. Adopting the probabilistic formulation with discrete levels not only offers all the advantages of the traditional VQ approach, but also compensates for the inherent hard decision effects (i.e., one-to-one mapping) by turning the solution into a soft-decision process.

\[ p(C_q | z_t) = \frac{p(C_q) p(z_t | C_q)}{\sum_{q=0}^{Q} p(C_q) p(z_t | C_q)} \]

4. EXPERIMENTS

The performance of the proposed probabilistic transformation is evaluated on 2 female and 2 male speakers (native speakers of American English) taken from the UTScope database [11]. Each speaker repeats 35 phonetically balanced sentence prompts and the speech signal is simultaneously recorded using CTM and PMIC.

In the experiments, Leave-one-out (LOO) cross-validation scheme is used such that out of the 35 sentence prompts, 34 sentences are used to construct the transformation matrices, while the remaining sentence is employed for evaluation. The final performance measure of the transformation method is obtained by averaging results across all evaluations. The LOO strategy is adopted to ensure that the results are not biased due to data partitioning.

The overall quality of the estimated CTM speech is assessed using five different objective metrics. The Perceptual evaluation of speech quality (PESQ) measure defines the difference between the loudness spectra averaged over time and frequency of a time aligned target and enhanced signal. The PESQ has been shown to exhibit a high correlation coefficient of \( \rho = 0.89 \) with the mean opinion score (MOS) [12]. Frequency weighted segmental SNR (fwsNRseg) is another measure highly correlated with the subjective listening tests [12]. The log-likelihood ratio (LLR) and Itakura-Saito distance (ISD) metrics measure the perceptual difference between the spectra of the target and the enhanced speech signals, while the weighted spectral slope (WSS) measure is designed to magnify the differences in the spectral peak locations between the two signals (for more details on these metrics see [12]). Since each metric is formulated based on a particular speech feature, it is clear that no single metric is optimal across all conditions [13]. Therefore, to have a reliable performance evaluation, the results are represented in terms of all the above noted metrics.

\[ p(C_q | z_t) = \frac{p(C_q) p(z_t | C_q)}{\sum_{q=0}^{Q} p(C_q) p(z_t | C_q)} \]

5. RESULTS

In this section, performance of the proposed approach is reported based on (i) spectral comparison (ii) five distinct objective measures. Fig. 3 shows wideband spectrograms of the estimated and original CTM signals for the same utterance as in Fig. 2. Comparing the two figures, it can be seen that the proposed approach is able to effectively reconstruct the attenuated spectral content and minimize the changes in formant locations and bandwidth. More specifically, the fricative sound /sh/ (100-200 ms interval), which is attenuated in the PMIC signal (see Fig. 2), is effectively enhanced. Moreover, the frequency band 1-3 kHz, which is almost nonexistent in the PMIC spectrum due to the alternative pickup location, is reliably estimated. The formant locations and bandwidth changes from the PMIC spectrum as compared to that of the CTM, are adjusted as well. The formant transitions in the 1000-1400 ms range are also strengthened by applying the proposed transformation.

To objectively assess the quality of the estimated signals, five distinct metrics are employed. The VQ codebook size is varied from 2 to 8 bits and the best performance, which is obtained by averaging the results across all the LOO cross-validation evaluations, is reported for each metric. Table 1 shows the performance of the proposed probabilistic transformation (PT) based on the PESQ, fwsNRseg, ISD, LLR, and WSS measures. The PESQ represents the perceptual clarity in the speech quality and the fwsNRseg represents the improvement in the segmental SNR, hence higher scores on these measures indicate better performance. The other three measures indicate the spectral closeness of the transformed signal to the target signal (CTM), therefore lower values on these represent better performance. The values are in terms of mean and standard deviation of the evaluations to better illustrate the reliability of the approach. Baseline performance is computed by comparing the original PMIC and CTM signals. The traditional plain VQ algorithm is also implemented to provide a benchmark for highlighting the advantages of incorporating the soft-decision strategy versus the inherent hard-decision existing within the VQ framework. Moreover, to demonstrate the importance of gain factor estimation, the results are given for two distinct cases, with (VQ-G and PT-G) and without gain estimation (VQ and PT). Two important observations are made during the experiments. First, the best performance for the plain VQ method is generally obtained when the codebook size is greater than or equal to 7 bits (7 and 8 bits), while the best performance for the proposed approach is generally obtained when the codebook size is 6 bits. Second, for the plain VQ, increasing the codebook size beyond 8 bits results in no improvement or aggravates the performance. Increasing the codebook size beyond 6 bits for the probabilistic transformation gives rise to some audible glitches in the estimated signals. This is due to speech data scarcity (34 sentences each between 2 to 5 seconds) in some VQ clusters, making the autocorrelation matrix \( R_q \) ill-conditioned, caus-
ing the estimated vocal filter to be unstable. As can be seen from the table: (i) under all the scenarios and across all the objective quality measures, the plain VQ and the proposed approach improve the speech quality and intelligibility, (ii) the gain estimation boosts the intelligibility measures, the plain VQ and the proposed approach improve the speech quality and intelligibility, (iii) the proposed approach, with or without gain estimation, performs far better than the hard-decision based plain VQ method.

6. CONCLUSION

A probabilistic transformation approach for improving overall quality and intelligibility of the noise-robust physiological microphone speech was presented. Objective assessment of the estimated signals with five distinct metrics indicated that this approach is quite effective at removing the metallic and muffled nature prevalent in PMIC signals. The PMIC is a vital tool for human-to-human speech communication under adverse noisy conditions and the proposed method can increase its functionality even further. Another application that could benefit from our approach could be patients treated for vocal fold pathologies, and those individuals using artificial larynxes. Future work involves improving the proposed system by including other acoustic features in the conditioning feature vector as well as providing alternative modeling for the excitation signal.

7. REFERENCES


---

Table 1. Objective assessment of overall quality of the speech signal estimated using the proposed probabilistic transformation (PT) and the plain VQ method, with (VQ-G and PT-G) and without (VQ and PT) gain estimation, based on five distinct quality metrics. Best performing metrics are in bold, with mean ± std-dev. M and F denote male and female speaker, respectively.

<table>
<thead>
<tr>
<th>SPKR</th>
<th>Method</th>
<th>PESQ</th>
<th>fSNRseg</th>
<th>ISD</th>
<th>LLR</th>
<th>WSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>VQ</td>
<td>1.84 ± 0.18</td>
<td>6.30 ± 0.60</td>
<td>7.28 ± 2.27</td>
<td>1.37 ± 0.11</td>
<td>30.04 ± 2.21</td>
</tr>
<tr>
<td>M1</td>
<td>VQ</td>
<td>2.01 ± 0.27</td>
<td>9.28 ± 0.99</td>
<td>2.05 ± 0.42</td>
<td>0.81 ± 0.09</td>
<td>29.25 ± 2.30</td>
</tr>
<tr>
<td></td>
<td>VQ-G</td>
<td>2.05 ± 0.29</td>
<td>9.25 ± 0.80</td>
<td>2.07 ± 0.55</td>
<td>0.84 ± 0.09</td>
<td>29.68 ± 2.15</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>2.45 ± 0.16</td>
<td>10.73 ± 0.99</td>
<td>1.31 ± 0.18</td>
<td>0.50 ± 0.08</td>
<td>24.33 ± 1.90</td>
</tr>
<tr>
<td></td>
<td>PT-G</td>
<td>2.58 ± 0.22</td>
<td>10.84 ± 1.01</td>
<td>1.03 ± 0.20</td>
<td>0.49 ± 0.08</td>
<td>24.10 ± 1.98</td>
</tr>
<tr>
<td>Baseline</td>
<td>VQ</td>
<td>1.97 ± 0.19</td>
<td>7.67 ± 0.47</td>
<td>2.29 ± 0.6</td>
<td>1.11 ± 0.08</td>
<td>28.71 ± 1.83</td>
</tr>
<tr>
<td>M2</td>
<td>VQ</td>
<td>2.25 ± 0.20</td>
<td>9.41 ± 1.02</td>
<td>3.47 ± 1.95</td>
<td>0.62 ± 0.08</td>
<td>27.70 ± 2.51</td>
</tr>
<tr>
<td></td>
<td>VQ-G</td>
<td>2.34 ± 0.19</td>
<td>9.26 ± 0.90</td>
<td>1.63 ± 0.49</td>
<td>0.63 ± 0.07</td>
<td>27.92 ± 2.34</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>2.64 ± 0.16</td>
<td>10.12 ± 0.64</td>
<td>3.43 ± 2.12</td>
<td>0.45 ± 0.05</td>
<td>24.83 ± 2.23</td>
</tr>
<tr>
<td></td>
<td>PT-G</td>
<td>2.70 ± 0.14</td>
<td>10.10 ± 0.61</td>
<td>0.92 ± 0.12</td>
<td>0.45 ± 0.05</td>
<td>25.14 ± 2.30</td>
</tr>
<tr>
<td>Baseline</td>
<td>VQ</td>
<td>1.72 ± 0.11</td>
<td>5.86 ± 0.59</td>
<td>2.33 ± 0.23</td>
<td>1.43 ± 0.12</td>
<td>33.38 ± 2.29</td>
</tr>
<tr>
<td>F1</td>
<td>VQ</td>
<td>2.07 ± 0.13</td>
<td>8.05 ± 0.51</td>
<td>4.45 ± 1.61</td>
<td>0.85 ± 0.12</td>
<td>31.32 ± 1.99</td>
</tr>
<tr>
<td></td>
<td>VQ-G</td>
<td>2.10 ± 0.13</td>
<td>8.00 ± 0.49</td>
<td>2.28 ± 0.78</td>
<td>0.87 ± 0.11</td>
<td>31.54 ± 2.03</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>2.42 ± 0.15</td>
<td>8.73 ± 0.49</td>
<td>3.47 ± 1.48</td>
<td>0.67 ± 0.12</td>
<td>28.20 ± 2.02</td>
</tr>
<tr>
<td></td>
<td>PT-G</td>
<td>2.51 ± 0.16</td>
<td>8.67 ± 0.51</td>
<td>1.44 ± 0.42</td>
<td>0.67 ± 0.11</td>
<td>28.33 ± 2.02</td>
</tr>
<tr>
<td>Baseline</td>
<td>VQ</td>
<td>1.99 ± 0.13</td>
<td>5.63 ± 0.59</td>
<td>2.31 ± 0.37</td>
<td>1.19 ± 0.09</td>
<td>30.60 ± 1.36</td>
</tr>
<tr>
<td>F2</td>
<td>VQ</td>
<td>2.27 ± 0.16</td>
<td>8.66 ± 0.46</td>
<td>4.34 ± 1.45</td>
<td>0.58 ± 0.07</td>
<td>29.86 ± 1.65</td>
</tr>
<tr>
<td></td>
<td>VQ-G</td>
<td>2.33 ± 0.21</td>
<td>8.77 ± 0.43</td>
<td>1.41 ± 0.65</td>
<td>0.56 ± 0.07</td>
<td>29.21 ± 1.35</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>2.56 ± 0.14</td>
<td>9.25 ± 0.46</td>
<td>4.23 ± 2.22</td>
<td>0.46 ± 0.05</td>
<td>27.56 ± 1.53</td>
</tr>
<tr>
<td></td>
<td>PT-G</td>
<td>2.61 ± 0.14</td>
<td>9.36 ± 0.34</td>
<td>0.90 ± 0.18</td>
<td>0.45 ± 0.05</td>
<td>27.05 ± 1.57</td>
</tr>
</tbody>
</table>