# Source and Filter Estimation for Throat-Microphone Speech Enhancement

M. A. Tuğtekin Turan, *Student Member, IEEE*, and Engin Erzin, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new statistical enhancement system for throat microphone recordings through source and filter separation. Throat microphones (TM) are skin-attached piezoelectric sensors that can capture speech sound signals in the form of tissue vibrations. Due to their limited bandwidth, TM recorded speech suffers from intelligibility and naturalness. In this paper, we investigate learning phone-dependent Gaussian mixture model (GMM)-based statistical mappings using parallel recordings of acoustic microphone (AM) and TM for enhancement of the spectral envelope and excitation signals of the TM speech. The proposed mappings address the phone-dependent variability of tissue conduction with TM recordings. While the spectral envelope mapping estimates the line spectral frequency (LSF) representation of AM from TM recordings, the excitation mapping is constructed based on the spectral energy difference (SED) of AM and TM excitation signals. The excitation enhancement is modeled as an estimation of the SED features from the TM signal. The proposed enhancement system is evaluated using both objective and subjective tests. Objective evaluations are performed with the log-spectral distortion (LSD), the wideband perceptual evaluation of speech quality (PESQ) and mean-squared error (MSE) metrics. Subjective evaluations are performed with an A/B comparison test. Experimental results indicate that the proposed phone-dependent mappings exhibit enhancements over phone-independent mappings. Furthermore enhancement of the TM excitation through statistical mappings of the SED features introduces significant objective and subjective performance improvements to the enhancement of TM recordings.**

***Index Terms*—Speech enhancement, throat microphone, Gaussian mixture model, statistical mapping.**

## I. Introduction

NON-ACOUSTIC sensor configurations have been increasingly studied in the recent literature for the speech enhancement problem to deliver robust speech processing applications. Environmental conditions, such as a presence of any background noise or wind turbulence, motivated researchers to use mediums other than the acoustic pathway to capture robust speech signal representations. Human tissue, infrared ray, light wave and laser are among the non-acoustic mediums to capture speech signals. Although sensors are developed for these non-acoustic mediums, their applications are limited due to their poor signal representation capabilities and/or their uncomfortable mounting schemes. Recently, there is an increasing trend in research and development of wearable devices that will also deliver ubiquitous, affordable and usable non-acoustic sensors in the future.

State-of-the-art non-acoustic sensor technologies include electroglottograph (EGG), glottal electromagnetic micro-power sensor (GEMS), non-audible murmur (NAM), bone-conducting (BCM) and throat (TM) microphones [1]–[5]. The EGG, which is also referred as laryngograph, has been designed to measure vibratory characteristics of the vocal folds. Weak electrical current on movements of the vocal folds was discussed first by Fabre [1] where he called this method as high-frequency glottography. The EGG principally provides a waveform representation of vocal fold dynamics and relative contact patterns during phonation. The GEMS, which is a device developed by Aliph Corporation, transmits low-power electro-magnetic (EM) waves to the glottis and the reflected signal captures tissue movements of voiced speech including opening and closing phases of the glottis via a small antenna located on the throat [2]. The NAM, which is developed by Nakajima *et al.* [3], is a typical contact-microphone whose sensor is based on a medical stethoscope used for monitoring internal sounds of the human body. It is attached behind the speaker's ear and records quietly uttered speech that cannot be captured acoustically through the tissue contact. The BCM captures speech via bone and tissues near the speaker's ear. It converts electric signals into mechanical vibrations and captures sound from the internal ear through the cranial bones. Its first description was depicted by Gernsback's patent in 1923 [4]. The TM, which has been used in military applications and radio communication for several decades, can capture speech signals in the form of vibrations and resonances of vocal cords through skin-attached piezoelectric sensors. It is also used for patients who have lost their voices due to injury or illness, or patients who have temporary speech loss after a tracheotomy [5]. The TM as a piezoelectric transducer can pick up a speech signal by absorbing the vibrations generated from the speech production system. Thus, it is robust to environmental acoustic conditions. However, it can only capture very low basebands of speech signals since tissue and bones act as a low-pass filter, which greatly reduces the intelligibility of the recorded speech, due to the limited frequency bandwidth.

We can present speech processing research through non-acoustic sensors (NAS) in two categories. The first line of research takes non-acoustic sensors as complementary information sources in addition to acoustic microphone (AM) recordings. In this line of research joint processing of NAS and

Manuscript received August 21, 2015; revised October 23, 2015; accepted October 27, 2015. Date of publication November 09, 2015; date of current version January 05, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

The authors are with the College of Engineering, Koc University, Istanbul 34450, Turkey (e-mail: mturan@ku.edu.tr; eerzin@ku.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2499040

AM recordings have been studied for robust speech recognition, speech enhancement and speech coding problems. The second line of research investigates NAS as the primary source of information in the absence of AM recordings, in which speech enhancement appears as the main research problem.

Recent literature includes several interesting studies for the first line of research on joint processing of NAS and AM recordings. In one of the early studies, Viswanathan *et al.* presented a two sensor system involving an accelerometer mounted on the speaker's throat and a noise-canceling microphone located close to the lips [6]. Close-talking first- and second-order differential microphones are designed to be placed close to the lips where the sound field has a large spatial gradient and the frequency response of the microphone is flat. Second-order differential microphones using a single element piezoelectric transducer have been suggested for use in a very noisy environment of aircraft communication systems to enhance a noisy signal for improved speech recognition.

Graciarena *et al.* proposed a pioneering work that estimates clean acoustic speech features using the probabilistic optimum filter (POF) mapping with combined TM and AM recordings [7]. The POF mapping is a piecewise linear transformation applied to noisy feature space to estimate the clean space [8]. It maps the temporal sequence of noisy Mel-cepstral features from the AM and TM recordings, which results in an optimal combination of the noisy acoustic and robust throat speech.

In [9], Zheng *et al.* detect whether the speaker is talking by combining the two channels from AM and BCM recordings, where the active speaker detection eliminates more than 90% of background speech. In the same framework, they tried to suppress non-stationary noises in both automatic speech recognition and enhancement tasks. They also developed a SPLICE-based mapping scheme to estimate clean speech features from BCM recordings in [10]. One of the problems, associated with the bone sensor signal in noisy environments, is teeth-clacks which are caused when the user's upper and lower jaws unconsciously come in contact with each other. Subramanya *et al.* [11] proposed an algorithm to remove this leakage by estimating the transfer function between the two sensors during regions of non-speech activity. They reported a method to extract formant information from the bone-sensor recordings through synthesizing speech waveforms based on the LPC cepstra. The Expectation-Maximization algorithm was used for parameter estimation from the noisy speech in the combination of AM and BCM recordings for robust speech recognition in [12].

The NAM microphone, which captures non-audible murmur, is mainly used for privacy purposes while communicating with speech recognition engines. On the other hand, it can be useful for people who have physical difficulties in speech production [3]. Heracleous *et al.* [13] investigated NAM recognition in noisy environments and the effect of the Lombard reflex on speech recognition. They also proposed a method based on multi-level Lombard hidden Markov models (HMM) to recognize arbitrary Lombard NAM utterances. In [14], a new hardware prototype that integrates several heterogeneous sensors such as bone, throat and in-ear microphones into a single headset has been presented. This prototype was used for robust speech detection in noisy environments, especially in non-stationary noise.

In another multi-sensory study, speech recorded from throat and acoustic channels is processed by parallel speech recognition systems and later a decision fusion yields robust speech recognition to background noise [15]. Adaptation methods, including maximum likelihood linear regression, sigmoidal low-pass filtering and linear multivariate regression, for recognition of soft whispers captured with a TM, were presented in [16]. TM signals were used for voice activity detection (VAD) to improve speech recognizer performance in [17]. It was reported that recognition accuracies in non-stationary noise improve significantly compared to when VAD is executed on a conventional microphone signal. A framework that defines a temporal correlation model between simultaneously recorded TM and AM speech was developed in [18]. This framework aims to learn joint sub-phone patterns of the TM and AM recordings that define temporally correlated neighborhoods through a parallel branch hidden Markov model (HMM) structure. The resulting temporal correlation model is employed to estimate acoustic features, which are spectrally richer, from throat counterparts through linear prediction analysis. The TM and estimated AM features are then used in a multi-modal speech recognition system.

Non-acoustic sensors can preserve speech attributes that are lost in the noisy acoustic signal, such as low-energy consonant voice bars, nasality, and glottal excitation. Quatieri *et al.* investigate methods of fusing non-acoustic low-frequency and pitch content with acoustic-microphone for low-rate coding of speech [19]. Low-rate coding paradigms involving this multi-band fusion approach and speaker-dependent source characterization that exploit non-acoustic sensor outputs under high-noise environments has been also considered by the same group in [20].

In the second line of research enhancement of NAS recordings has been studied. In [21], spectral and excitation features of acoustic speech are estimated from the spectral features 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 was reported to suffer from unnatural prosody due to the synthetic fundamental frequency generation. In another study [22], the transfer characteristics of BCM and AM 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. Since the transfer function of the bone-conduction path is speaker and microphone dependent, the transfer function should be individualized for effective equalization. Speaker-dependent short-term FFT based equalization is proposed using simultaneously recorded AM and BCM speech.

Other issues such as detection of body-internal sounds like swallowing, chewing or whispering have also appeared and been discussed in some research. Swallowing sound signals are collected by TM in [23], [24] to achieve accurate methods for monitoring ingestive behavior. Methods based on the Mel-scale Fourier spectrum, wavelet packets, and support vector machines (SVM) are investigated to reveal the effects

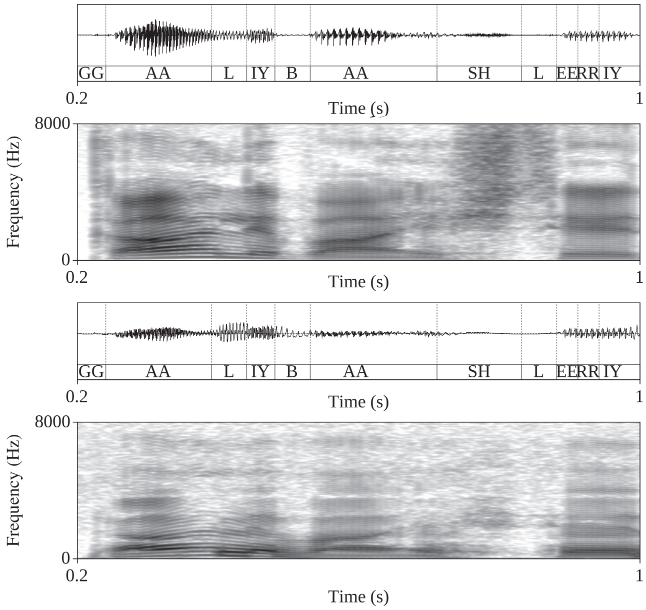

Fig. 1. A sample of parallel AM (top) and TM (bottom) recordings and their spectrograms with phonetic transcription of the Turkish utterance “galiba şileri”.

of epoch size and lagging on classification accuracy. Then, it is emphasized that their proposed methods can separate swallowing sounds from artifacts that originate from respiration, intrinsic speech, head movements, food ingestion, and ambient noise. Monitoring the swallowing using TM can also be used to analyze food intake behavior of individuals [25]. This kind of research has a potential to detect and analyze obesity through automated ingestion monitoring.

In this paper, our primary interest is the enhancement of TM speech recordings in the absence of AM speech. Although TM recordings are partly intelligible, the main problem arises from listening effort. TM recordings are muffled due to the low-pass characteristic of tissue conduction, note that this low-pass characteristic is also nonlinear due to non-homogeneous tissue structures. However, TM recordings capture pitch and some partial formant structure, and TM systems generally preferred under high ambient noise where the conventional microphones cannot be used. Fig. 1 shows sample waveforms and spectrograms of simultaneously recorded AM and TM recordings. We observe the low-pass characteristic of tissue propagation around 3 kHz cutoff frequency. In order to understand the perceptual difference between TM and AM speech, it is necessary to understand the vocal tract characteristics of TM recordings. Phones such as */sh/* and */l/*, which are realized over the friction of narrow-stream turbulent air with high-frequency spectral energy components, cannot preserve their spectral structures. On the other hand, phones such as */gg/* and */b/*, which are articulated via blocking the airflow with tongue or lips, have similar tendency in both AM and TM due to their limited bandwidth.

Enhancement of the TM speech is in certain aspects similar to the bandwidth extension problem of band-limited telephony speech. Artificial bandwidth extension (ABE) studies try to recover the missing high-frequency components of telephony speech [26], [27]. Hence, ABE approaches can avail valuable insights for the TM speech enhancement problem. One widely used framework for the ABE problem is splitting the telephone-band speech signal through source-filter separation. Then the source (excitation) and the filter (spectral envelope) of the narrow-band speech signal are extended separately and recombined to synthesize a wideband speech signal [28]. In the bandwidth extension framework, an extension of the excitation signal has been performed by modulation, which attains spectral continuation and a matching harmonic structure of the baseband [26]. In a sense, this method guarantees that the harmonics in the extended frequency band always match the harmonic structure of the baseband. However, this is expected to be a poor extension for the TM excitation signal since TM recordings are captured through tissue conduction with a nonlinear low-pass filtering. On the other hand, extension of spectral envelope, equivalently spectral mapping, has been studied widely for both ABE [26], [27] and speech conversion [29]. Jax and Vary introduced a hidden Markov model (HMM)-based wideband spectral envelope estimator [26]. Their HMM-based estimator can be modeled as a weighted sum of all estimations in all states with a soft mapping, which is defined by the emission probabilities of the narrowband observations and the state transition probabilities. Later Yagli *et al.* [27] modified this soft HMM-based mapping by decoding an optimal Viterbi path based on the temporal contour of the narrowband spectral envelope and then performing the minimum mean square error (MMSE) estimation of the wideband spectral envelope on this path. Stylianou *et al.* [29] presented one of the early works on continuous probabilistic mapping of the spectral envelope for the voice conversion problem, which is defined as modifying the speech signal of one speaker (source) so that it sounds like be pronounced by a different speaker (target). Their contribution includes the design of a new methodology for representing the relationship between two sets of spectral envelopes. Their proposed method is based on the Gaussian mixture model (GMM) of source speaker spectral envelopes. The conversion itself is represented by a continuous parametric function which takes into account the probabilistic classification provided by the mixture model. Later Toda *et al.* [30] improved the continuous probabilistic mapping by incorporating not only static but also dynamic feature statistics for the estimation of a spectral parameter trajectory.

In this paper, we present a complete system for enhancement of TM recordings through source and filter separation. In this enhancement system we address proper mappings of both excitation and spectral envelope to synthesize the perceptually improved speech signal from TM recordings. Main contributions of the proposed enhancement system are: i) mappings of both excitation and spectral envelope are phone-dependent to address the phone-dependent variability of tissue conduction with TM recordings; ii) probabilistic mapping structures and their fusion are investigated for the excitation enhancement; and iii) objective and subjective quality improvements are reported for phone-dependent enhancement of both excitation and spectral envelope. The proposed probabilistic mapping differs from the state-of-the-art mapping techniques of [22], [26], [27], [31] by addressing phone-dependent context modeling for TM recordings. In comparison to our recent works in [32], [33], in this paper we extensively investigate excitation enhancement techniques and evaluate the complete TM enhancement system through objective and subjective metrics.

The remainder of paper is organized as follows: the proposed TM speech enhancement system is given in Section II. Experimental evaluations and results are addressed in Section III. Finally, Section IV includes the discussions and future research directions.

## II. TM Enhancement System

### A. System Overview

In this section we present the proposed enhancement system of TM recordings through source and filter separation. In this system, enhancement of TM recordings is formalized as probabilistic mappings of spectral envelope and excitation representations to the representations of AM recordings. The enhanced spectral envelope and excitation of TM are then used to synthesize perceptually improved speech signal. First, let us define the source-filter separation through the linear prediction (LP) filter model of time-aligned TM and AM speech as

$$\begin{aligned} S_T(z) &= E_T(z)H_T(z) \\ S_A(z) &= E_A(z)H_A(z), \end{aligned} \tag{1}$$

where $H_T(z)$ and $H_A(z)$ are the linear prediction filters, and $E_T(z)$ and $E_A(z)$ are the excitation spectra of the TM and AM speech, respectively. In this paper we refer to elements of these representations as column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively representing the TM speech representation as an observable source $\mathcal{X}$ and the AM speech representation as a hidden source $\mathcal{Y}$.

A block diagram of the proposed enhancement system, which includes learning and enhancement parts, is given in Fig. 2. In the learning part, spectral envelope and excitation representations from time-aligned TM and AM speech are used to construct probabilistic mappings from the observable source (TM speech) to the hidden source (AM speech). Then in the enhancement part, hidden source (AM speech) representations are estimated from the observable source (TM speech) using the probabilistic mapping structures that are constructed in the learning part for spectral envelope and excitation representations. The enhanced TM speech is reconstructed through LP synthesis using the estimated hidden source representations.

### B. GMM-Based Probabilistic Mapping

The Gaussian mixture model (GMM) is a classic parametric model used in many pattern recognition techniques to represent multivariate probability distributions. GMM states that any general distribution of $x$ can be approximated by a sum of weighted Gaussian distributions:

$$P(\boldsymbol{x}) = \sum_{l=1}^{L} \omega_l \mathcal{N}_l(\boldsymbol{x}; \boldsymbol{\mu}_l, \boldsymbol{C}_l), \tag{2}$$

where $\mathcal{N}_l(\boldsymbol{x}; \boldsymbol{\mu}_l, \boldsymbol{C}_l)$ is the multi-variate Gaussian distribution with mean vector $\boldsymbol{\mu}_l$ and covariance matrix $\boldsymbol{C}_l$, $\omega_l$ is the mixture weight corresponding to the $l$-th mixture density and satisfies $\sum_{l=1}^{L} \omega_l = 1$ with $\omega_l \geq 0$.

In general GMM partitions the data space into $L$ clusters. In each cluster, data is modeled with a multivariate Gaussian distribution $\mathcal{N}_l(\boldsymbol{x}; \boldsymbol{\mu}_l, \boldsymbol{C}_l)$. Then the weighted sum of the cluster distributions forms the GMM distribution for the data space. In GMM covariance matrices are typically assumed to be diagonal for modeling and computational benefits. The GMM parameters can be estimated by the expectation-maximization algorithm [34]. Furthermore an initial vector quantization of the data space through the generalized Lloyd algorithm [35] may accelerate convergence.

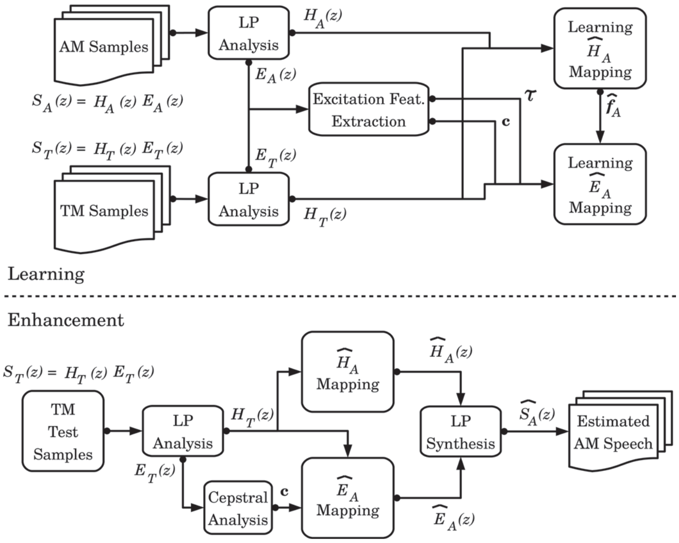

Fig. 2. Block diagram of the proposed enhancement system.

In defining GMM-based probabilistic mapping we can consider the observable source $\mathcal{X}$ and the hidden source $\mathcal{Y}$. When we assume that $\mathcal{X}$ and $\mathcal{Y}$ are jointly Gaussian, then the mean square estimation of a hidden source within a single data cluster, i.e., for an isolated Gaussian mixture, is defined as

$$\begin{aligned} \widetilde{\boldsymbol{y}}_l(\boldsymbol{x}) &= \mathcal{E}(\boldsymbol{y}_l | \boldsymbol{x}) \\ &= \boldsymbol{\mu}_{y,l} + \boldsymbol{C}_{yx,l}(\boldsymbol{C}_{xx,l})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{x,l}), \end{aligned} \tag{3}$$

where $\boldsymbol{C}_{yx,l}$ is the cross-covariance matrix between the observable and hidden sources for the $l$-th mixture and $\mathcal{E}()$ is the expectation operator. Then the GMM-based minimum mean square error (MMSE) estimation of the hidden source as defined in [29] and [36] is given as

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{l=1}^{L} P(\gamma_l | \boldsymbol{x}) \widetilde{\boldsymbol{y}}_l(\boldsymbol{x}), \tag{4}$$

where $\gamma_l$ is the $l$-th Gaussian mixture and $L$ represents the total number of Gaussian mixtures. The probability of the $l$-th Gaussian mixture given the observation $\boldsymbol{x}$ is defined as, the normalized Gaussian density function as

$$P(\gamma_l | \boldsymbol{x}) = \frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x,l}, \boldsymbol{C}_{xx,l})}{\sum_{i=1}^{L} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x,i}, \boldsymbol{C}_{xx,i})}. \tag{5}$$

In [32] we investigate two level partitioning of the observable data for the GMM distribution

$$P(\boldsymbol{x}) = \sum_{p=1}^{N} \sum_{l=1}^{L_p} \omega_{lp} \mathcal{N}_{lp}(\boldsymbol{x}; \boldsymbol{\mu}_{lp}, \boldsymbol{C}_{lp}), \tag{6}$$

where partitions over $p$ can define phonetic clusters of the acoustic data. Such a first level partitioning has been observed to improve spectral envelope mapping. Given the first level partitioning, or equivalently phonetic class, we can define a phone-dependent MMSE estimation of the hidden source as

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}|p) = \sum_{l=1}^{L_p} P(\gamma_l^p|\boldsymbol{x})\widetilde{\boldsymbol{y}}_l(\boldsymbol{x}|p), \tag{7}$$

where $\widetilde{\boldsymbol{y}}_l(\boldsymbol{x}|p)$ is the MMSE estimate of the hidden source within mixture $l$ for the given phone $p$:

$$\widetilde{\boldsymbol{y}}_l(\boldsymbol{x}|p) = \mathcal{E}(\boldsymbol{y}_l|\boldsymbol{x} \in p). \tag{8}$$

The total number of mixture components in the phone-independent mapping $\widehat{\boldsymbol{y}}(\boldsymbol{x})$ and the phone-dependent mapping $\widehat{\boldsymbol{y}}(\boldsymbol{x}|p)$ should be similar for a fair comparison. Hence we adjust the number of mixtures for the phone-dependent mapping as

$$L_p = 2^{\lfloor \log_2(r_p L) + 0.5 \rfloor}, \tag{9}$$

where $r_p$ represents the relative frequency of the phone $p$, $L_p$ is the number of mixtures for phone $p$, and $L$ is the total number of mixtures in the phone-independent mapping.

Note that $\widehat{\boldsymbol{y}}(\boldsymbol{x}|p)$ in (7) is a soft mapping. Alternatively we can define a hard mapping to choose the most likely MMSE estimation among all mixtures as

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}|p, l^x) = \widetilde{\boldsymbol{y}}_{l^x}(\boldsymbol{x}|p), \tag{10}$$

where $l^x$ is the index of the most likely mixture

$$l^x = \arg\max_l P(\gamma_l^p|\boldsymbol{x}). \tag{11}$$

Furthermore we define a fusion mapping using the hard mappings, which are available for several different representations of the observable data. Let us define $N$ different representations of the observable data as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. Then we can construct the fusion mapping of these $N$ representations using hard mappings over the observable data as

$$\begin{aligned} \widehat{y}_F(\boldsymbol{x}_1; \ldots; \boldsymbol{x}_N) &= \widehat{y}(\boldsymbol{x}_1|p, l^1) \oplus \cdots \oplus \widehat{y}(\boldsymbol{x}_N|p, l^N), \\ &= \widehat{y}(\boldsymbol{x}_n|p, l^n) \\ \text{for } n &= \arg\max_{i=1,\ldots,N} P(\gamma_{l^i}^p|\boldsymbol{x}_i). \end{aligned} \tag{12}$$

### C. Spectral Envelope Mapping

Spectral envelope is represented with the all-pole LP filter $H(z)$ as defined in (1). The line spectrum frequency (LSF) representation of the linear prediction filter is used to model the spectral envelope. We define frame level LSF representation vectors $\boldsymbol{f}_T$ and $\boldsymbol{f}_A$, respectively for the TM and AM sources. Then the observable and hidden sources of the spectral envelope can be respectively stated as

$$\mathcal{X}^H : \boldsymbol{f}_T \qquad \text{and} \qquad \mathcal{Y}^H : \boldsymbol{f}_A. \tag{13}$$

In [32] we investigate phone-dependent mappings for spectral envelope enhancement and observed significant improvements when the true phone-context is available to the spectral mapping. Hence we consider the phone-independent and phone-dependent mappings for enhancement of the spectral envelope. The estimated spectral envelope filter for the phone-independent mapping can be denoted as

$$\widehat{H}_A^{H_T} \leftrightarrow \widehat{\boldsymbol{f}}_A(\boldsymbol{f}_T), \tag{14}$$

where $\widehat{\boldsymbol{f}}_A(\boldsymbol{f}_T)$ is the phone-independent mapping as defined in (4). Similarly, the estimated spectral envelope filter for the phone-dependent mapping in (7) is denoted as

$$\widehat{H}_A^{H_T|p} \leftrightarrow \widehat{\boldsymbol{f}}_A(\boldsymbol{f}_T|p), \tag{15}$$

where $p$ represents the likely phone context, which is decoded by an HMM-based phoneme recognition system over the observable TM source. We also consider the true phone context, $p^*$, which is extracted by the force alignment procedure, in the learning and enhancement parts for comparative evaluations.

### D. Excitation Mapping

The phone-dependent variability of tissue conduction creates frequency selective attenuations in the TM recordings. Although the spectral envelope enhancement compensates these attenuations using the linear prediction representation, phone and frequency dependent spectral differences still exist between the AM and TM excitation signals. The missing spectral details of the TM excitation are observed as an important source of degradation in the TM voice quality. Hence we model the missing spectral details as the mismatch between the AM and TM excitation signal that can be represented as a phone dependent spectral energy difference (SED) vector. Then we train a probabilistic mapping from the observable TM spectral features to the phone dependent SED representation.

Let us first define the spectral band energy vector $\boldsymbol{e}$ for the representation of the excitation spectra as

$$e(b) = \log\left[\sum_{k=1}^{K-1} w_b(k)|E(k)|^2\right] \quad \text{for } b = 1, \ldots, B, \tag{16}$$

where $B$ is the number of spectral bands, $E$ is the $2K$-point DFT of the excitation signal, and $w_b$ is the window function for spectral band $b$. The window function is defined as

$$w_b(k) = \begin{cases} 0 & k < f_{b-1} \; \mathit{or} \; k > f_{b+1}, \\ \frac{k - f_{b-1}}{f_b - f_{b-1}} & f_{b-1} \leq k \leq f_b, \\ \frac{f_{b+1} - k}{f_{b+1} - f_b} & f_b < k \leq f_{b+1}, \end{cases} \tag{17}$$

where $f_b$ is the $b$-th center frequency index and $w_b$ is the $b$-th triangular filter for $b = 0, 1, \ldots, B+1$ with $f_0 = 0$ and $f_{B+1} = K$, which are taken as boundary frequency indexes.

Then the spectral energy difference (SED) between AM and TM excitation signals is defined as

$$\tau_b = e_A(b) - e_T(b) \quad \text{for } b = 1, \ldots, B. \tag{18}$$

Note that the SED vector is defined as the hidden source for the excitation mapping

$$\mathcal{Y}^E : \tau. \tag{19}$$

We use the LSF feature set $\boldsymbol{f}_T$, the estimated LSF feature set $\widehat{\boldsymbol{f}}_A$ and excitation cepstrum $\boldsymbol{c}$ as the observable sources

$$\mathcal{X}^E : \boldsymbol{f}_T, \widehat{\boldsymbol{f}}_A, \boldsymbol{c}, \tag{20}$$

where the excitation cepstrum is defined as the discrete cosine transform of the spectral band energy of TM excitation

$$c(n) = \sum_{b=1}^{B} e_T(b) \cos\left(\pi n(b - 1/2)/B\right). \tag{21}$$

We construct phone-dependent estimators for the SED vector using only the TM spectra and only the excitation cepstrum respectively as the observable sources

$$\widehat{E}_A^{H_T|p} : \widehat{\tau}(\boldsymbol{f}_T|p), \tag{22}$$

$$\widehat{E}_A^{\boldsymbol{c}|p} : \widehat{\tau}(\boldsymbol{c}|p), \tag{23}$$

and with both the TM spectra and the excitation cepstrum as the observable sources

$$\widehat{E}_A^{\boldsymbol{c}H_T|p} : \widehat{\tau}(\boldsymbol{c}, \boldsymbol{f}_T|p). \tag{24}$$

Likewise in (14), it is possible to define a phone-independent mapping with the TM spectra and the excitation cepstrum as

$$\widehat{E}_A^{\boldsymbol{c}H_T} : \widehat{\tau}(\boldsymbol{c}, \boldsymbol{f}_T). \tag{25}$$

We also consider the enhanced TM spectra, that is the estimated AM spectral envelope representation $\widehat{\boldsymbol{f}}_A$, as a possible observable source, and construct the following SED estimator

$$\widehat{E}_A^{\boldsymbol{c}\widehat{H}_A|p} : \widehat{\tau}(\boldsymbol{c}, \widehat{\boldsymbol{f}}_A|p). \tag{26}$$

Finally we define a fusion mapping for the SED estimation over the above four observable sources as defined in (12)

$$\widehat{E}_A^{F|p} : \widehat{\tau}_F(\boldsymbol{c}; \boldsymbol{f}_T; \boldsymbol{c}, \boldsymbol{f}_T; \boldsymbol{c}, \widehat{\boldsymbol{f}}_A). \tag{27}$$

Once we get the estimated SED vector $\widehat{\tau}$, the enhanced TM excitation spectra can be extracted as

$$\widehat{E}_A(k) = \sum_{b=0}^{B+1} w_b(k) 10^{\widehat{\tau}_b} E_T(k) \quad k = 1, \ldots, K-1, \tag{28}$$

where the boundary SED values are taken as $\widehat{\tau}_0 = \widehat{\tau}_1$ and $\widehat{\tau}_{B+1} = \widehat{\tau}_B$. Furthermore, the enhanced TM excitation signal can be reconstructed by overlap-and-add of the inverse DTF of the enhanced excitation spectra $\widehat{E}_A$.

## III. Enhancement Experiments

We perform experiments on a synchronous TM and AM database which consists of 799 phonetically-balanced sentences from one male speaker at 16-kHz sampling rate. An IASUS-GP3 headset and Sony condenser tie-pin microphone are used for the TM and AM respectively. Experimental evaluations are performed through 10-fold cross validation. That is, we use 90% of the database in the learning phase and the remaining 10% of the database is used in the enhancement evaluations. This procedure is repeated ten times to cover all the database in the enhancement evaluations.

TABLE I
The Turkish METUBet Phonetic Alphabet with 8 Articulation Attributes

| **Rounded** | | **Stops** | | **Fricatives** | |
|---|---|---|---|---|---|
| AA | anı | B | bal | H | hasta |
| A | laf | D | dede | J | müjde |
| I | ısı | GG | karga | F | fasıl |
| E | elma | G | genç | S | ses |
| EE | dere | KK | akıl | SH | aşı |
| IY | simit | K | kedi | VV | var |
| **Unrounded** | | P | ip | V | tavuk |
| O | soru | T | ütü | Z | azık |
| U | kulak | **Liquids** | | ZH | yoz |
| OE | örtü | LL | kul | **Affiricates** | |
| UE | ümit | L | leylek | C | cam |
| **Nasals** | | RR | Irmak | CH | seçim |
| M | dam | RH | bir | **Glide** | |
| NN | ani | R | raf | Y | yat |
| N | süngü | | | | |

In this study, spectral envelope is represented with the line spectrum frequency (LSF) parametrization of the linear prediction filter and excitation spectra are extracted with the short time Fourier transform (STFT). The spectral representations are extracted as 16-th order linear prediction filters over 30 ms Hamming windows with 10 ms frame shifts. For the short time Fourier transform (STFT), we again use 30 ms Hamming analysis windows over 10 ms frames.

In the experimental evaluations we consider two different phone contexts: the true and likely phone contexts. The true phone context, $p^*$, is extracted by phonetic transcription and considered as the most informative upper bound for the phone-dependent models. The phonetic transcription is performed using the Turkish phonetic dictionary METUbet [37] and the phone level alignment is performed using forced-alignment and visual inspection. The METUbet phonetic alphabet is given in Table I, where phones are categorized into 8 different manner of articulations. The likely phone context, $p$, is decoded by an HMM-based phoneme recognition system over the observable source $\mathcal{X}$, that is, the TM database. The HMM-based phoneme recognition is performed with a 3-state and 256-mixture density phone level HMM recognizer, which is trained over recordings of 11 male speakers of the TM database in [18]. The average phone recognition performance is obtained as 62.22%.

### A. Objective Evaluations

Evaluations of the TM speech enhancement are performed with three distinct objective metrics: the logarithmic spectral distortion (LSD), the perceptual evaluation of wide-band speech quality (PESQ) and the mean-squared error (MSE). The LSD is a widely used metric for spectral envelope quality assessment. It is also symmetric, unlike the Itakura-Saito metric. The LSD metric assesses the quality of the estimated spectral envelope with respect to the original spectra, and is defined as,

$$d_{LSD} = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| 20 \log_{10}(H_A(\omega)) - 20 \log_{10}(\widehat{H}_A(\omega)) \right|^2 d\omega},$$

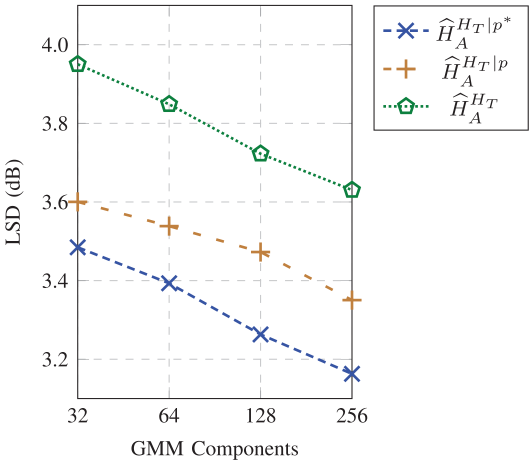

Fig. 3. Average LSD scores of the proposed spectral envelope mapping schemes.

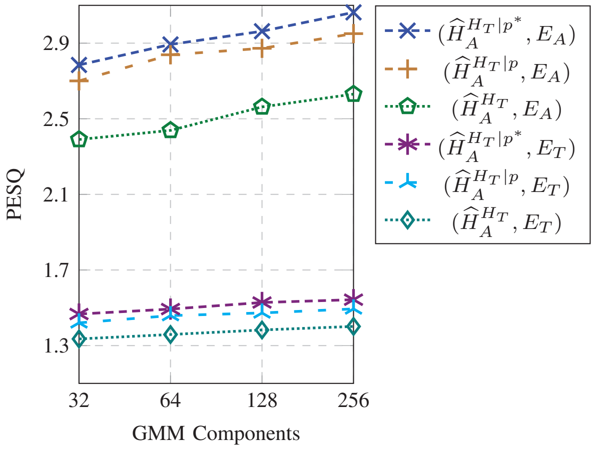

Fig. 4. Average PESQ scores of the proposed spectral envelope mapping schemes with the acoustic and throat excitation signals.

where $H_A(\omega)$ and $\widehat{H}_A(\omega)$ represent the original and estimated acoustic spectral envelopes, respectively. The ITU-T Standard PESQ [38] is employed as the second objective metric to evaluate the perceptual quality of the enhanced TM speech signal compared to the reference AM target. The PESQ algorithm predicts opinion scores of a degraded speech sample in (0, 5) range, where higher score indicates better quality. We also considered the MSE metric to evaluate estimated SED features since GMM mapping aims to minimize the mean-squared error in the estimation. The MSE metric for the SED features can be defined as,

$$d_{MSE} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\tau}_b - \tau_b)^2. \tag{29}$$

*1) Spectral Envelope Enhancement Experiments:* Fig. 3 presents the average LSD scores between the estimated filter $\widehat{H}_A(\omega)$ and the original acoustic filter $H_A(\omega)$. The best performing scheme is observed as the phone-dependent mapping when the true phone context is known, $\widehat{H}_A^{H_T|p^*}$, as defined in (15). On the other hand the phone-independent mapping $\widehat{H}_A^{H_T}$ has the lowest performance. The phone-dependent mapping $\widehat{H}_A^{H_T|p}$ with the likely phone context performs better than the phone-independent mapping but remains lower than the $\widehat{H}_A^{H_T|p^*}$ mapping, which is considered as the most informative upper bound for the phone-dependent models. In general we can argue that the phone-independent GMM mapping creates over-smoothing in the estimation and furthermore defining a phone context for the GMM mapping improves estimation performance.

We can synthesize an enhanced speech signal using the estimated spectral envelopes and acoustic or throat excitation signals. Fig. 4 presents average PESQ scores between the enhanced, $\widehat{S}_A$, and original, $S_A$, recordings. There are two important observations in these results: (i) Better excitation, $E_A$, delivers better PESQ scores, and (ii) the true, $p^*$, and likely, $p$, phone contexts perform better than the phone-independent mapping; their PESQ performances are close to each other and they deliver better improvement with the better excitation $E_A$.

TABLE II
Average PESQ Scores for All Possible Synthesis Scenarios of the TM and AM Representations

| Envelope | Excitation | PESQ (dB) |
|---|---|---|
| $H_T$ | $E_T$ | 1.25 |
| $H_A$ | $E_T$ | 1.62 |
| $H_T$ | $E_A$ | 2.14 |
| $H_A$ | $E_A$ | 4.46 |

*2) Excitation Enhancement Experiments:* We first investigate possible worst and best achievable case scenarios for the enhancement of the TM recordings when the AM speech is available. Table II presents average PESQ scores with the reference AM target for all possible synthesis scenarios of the TM and AM representations. As expected, the $(H_T, E_T)$ synthesis, equivalently TM recordings, and the $(H_A, E_A)$ synthesis, equivalently AM recordings, have respectively the lowest and highest average PESQ scores. On the other hand, the TM envelope and AM excitation, $(H_T, E_A)$, synthesis delivers higher PESQ score than the AM envelope and TM excitation, $(H_A, E_T)$, synthesis. This observation sets the importance of TM excitation enhancement for the perceived quality of the speech signal.

Note that in the processing of excitation signals, a 2048-point DFT is used over 30 ms Hamming-windowed excitation signals with a frame shift of 10 ms. The enhanced excitation signal, $\widehat{E}_A$, is reconstructed from the spectrum, $\widehat{E}_A(z)$, with inverse DFT and overlap-and-add schemes. In order to set the number of center frequencies, $B$, in (17), we evaluate average PESQ performance of the $(H_A, \widehat{E}_A^{\boldsymbol{C}H_T|p})$ synthesis for varying number of Mel and linear scale bands in Fig. 5. Although the average PESQ performances for the Mel and linear scale bands meet over $B = 24$ bands, the Mel scale has significant performance advantage under $B = 20$ number of bands. Hence we set $B = 16$ Mel scaled bands in our excitation enhancement evaluations.

As a second task in excitation enhancement we evaluate the MSE performances of the proposed SED estimation schemes. Fig. 6 presents the average MSE between estimated and original SED feature vectors. Note that the fusion mapping $\widehat{E}_A^{F|p}$

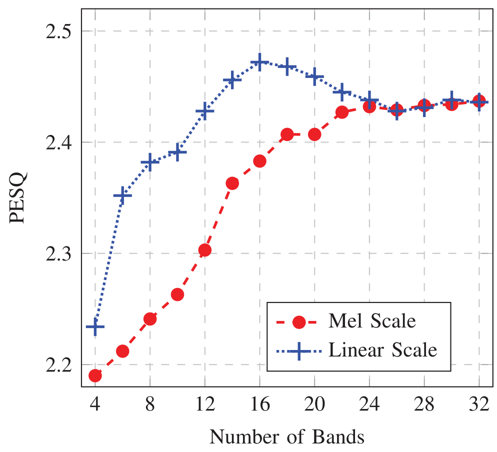

Fig. 5. Average PESQ performance of the $(H_A, \widehat{E}_A^{\mathbf{C}H_T|p})$ synthesis for varying number of Mel and linear scale bands.

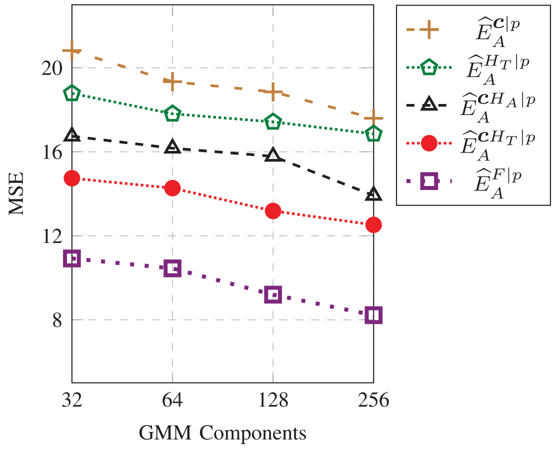

Fig. 6. Average MSE between estimated and original SED feature vectors for the proposed excitation mapping schemes.

has the lowest MSE. Furthermore the phone-dependent mapping with excitation cepstrum and TM spectra, $\widehat{E}_A^{\mathbf{C}H_T|p}$, does better than the phone-dependent mapping with excitation cepstrum and estimated AM spectra, $\widehat{E}_A^{\mathbf{C}H_A|p}$.

As a third task in excitation enhancement we evaluate the PESQ performance of the enhanced TM speech. In this evaluation spectral envelope enhancement is fixed to the $\widehat{H}_A^{H_T|p}$ mapping. In Fig. 7, average PESQ performances of the proposed excitation enhancement are presented for the true $p^*$ and likely $p$ phone contexts. Note that worst to best PESQ performance ordering for the mappings are same with the MSE performance ordering as in Fig. 6. The best performance is observed in the fusion mapping for both true and likely phone contexts in the $\widehat{E}_A^{F|p^*}$ and $\widehat{E}_A^{F|p}$ mappings, respectively. The true and likely phone contexts have similar tendencies; the fusion mappings bring significant PESQ improvements in both. Overall the excitation enhancement attains 30% average PESQ improvement from envelope only enhancement $(\widehat{H}_A^{H_T|p}, E_T)$ to envelope and excitation enhancement $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{F|p})$ of the TM speech.

The fourth task in excitation enhancement is to check PESQ performance of the enhanced speech in isolation of the proposed excitation enhancement schemes. For this purpose we set the spectral envelope from TM or AM recordings and evaluate contribution of the proposed excitation enhancement schemes to the perceived speech quality. The average PESQ performances are presented in Fig. 8 and they also deliver possible lower and upper bound PESQ improvements of the excitation enhancement. In this investigation we also consider phone-independent mapping $\widehat{E}_A^{\mathbf{C}H_T}$ for the SED feature, which is observed to perform significantly worse than the phone-dependent mappings. Also note that the fusion mapping $\widehat{E}_A^{F|p}$ introduces higher improvement with the better spectral envelope mapping $H_A$.

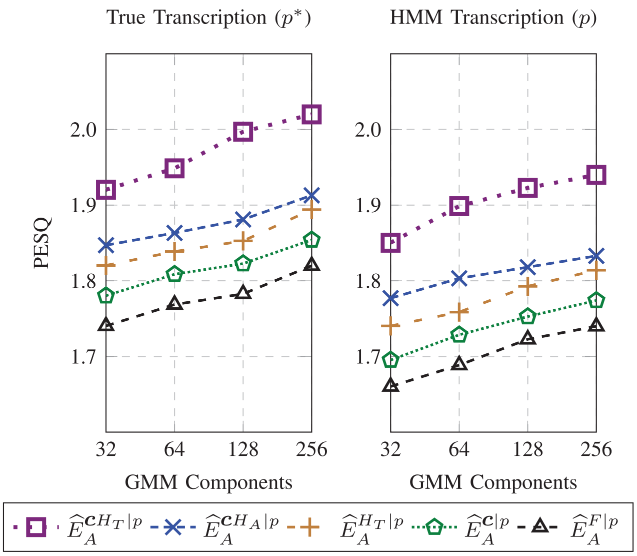

Fig. 7. Average PESQ performances of the proposed excitation mapping schemes with true and likely phone contexts when the envelope mapping is $\widehat{H}_A^{H_T|p}$.

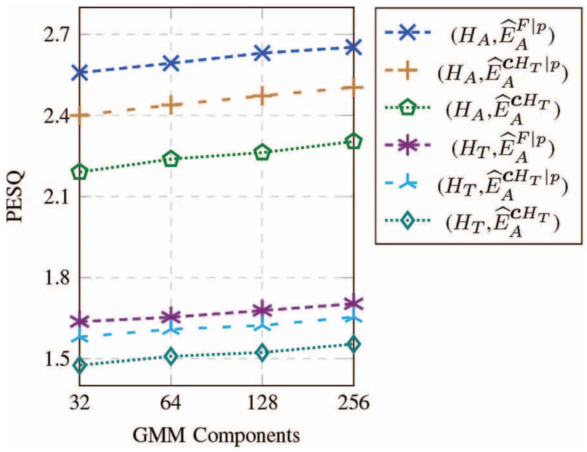

Fig. 8. Average PESQ performances of enhancement in isolation of the proposed excitation mapping schemes.

Finally we investigate performance of the spectral envelope and excitation mapping schemes on different phonetic attributes, including nasals, unrounded and rounded vowels, stops, liquids, fricatives, affricates and glides, which are reported in Table III. The spectral envelope mapping is evaluated with the LSD performance between $H_A$ and $\widehat{H}_A^{H_T|p}$. On the other hand, the excitation mapping is evaluated with the MSE performance between the original and synthesized SED vectors using the fusion mapping, $\widehat{E}_A^{F|p}$. As it was discussed earlier in the introduction, certain phonetic attributes are more robust with the TM and some others suffer more in terms of signal characterization and perceived quality. In Table III the smallest

TABLE III
AVERAGE LSD AND MSE PERFORMANCES AND OCCURRENCE FREQUENCIES OF DIFFERENT PHONETIC ATTRIBUTES

| MoA | LSD (dB) | MSE | Freq. (%) |
|---|---|---|---|
| Nasals | 4.74 | 1.49 | 9.37 |
| Unrounded Vowels | 5.99 | 3.52 | 8.95 |
| Stops | 6.21 | 8.37 | 19.67 |
| Liquids | 6.86 | 5.67 | 10.55 |
| Rounded Vowels | 7.19 | 5.73 | 32.13 |
| Glide | 7.30 | 11.21 | 2.87 |
| Fricatives | 9.51 | 16.68 | 13.40 |
| Affricates | 9.98 | 24.17 | 3.06 |

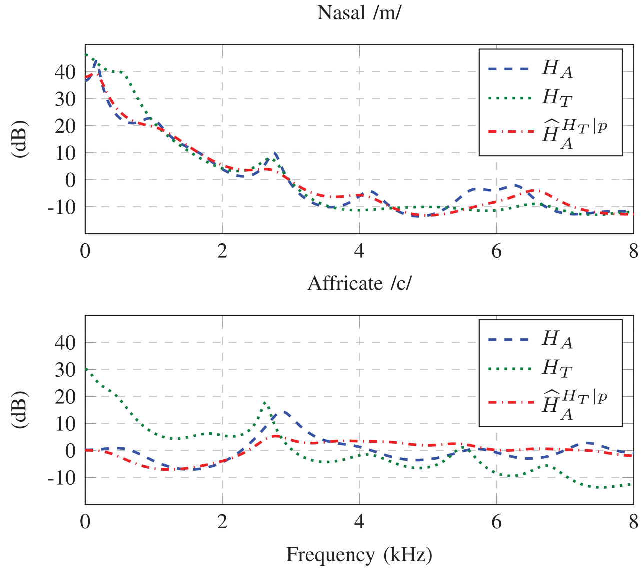

Fig. 9. Spectral envelope samples of nasal /m/ and affricate /c/ using the TM, AM and enhanced spectral envelope representations.

degradation is observed in nasal sounds, and affricates have the largest distortion. This characteristic was synchronously observed by the LSD and MSE performances for the envelope and excitation mappings, respectively. On the other hand, while liquids and rounded vowels suffer more with the envelope mapping, stops suffer more with the excitation mapping.

Fig. 9 presents sample envelope spectra for nasal /m/ and affricate /c/ sounds where AM and TM envelope spectra together with the estimated envelope spectra of the proposed phone-dependent mapping $\widehat{H}_A^{H_T|p}$ are presented. Note that the estimated spectra match closely to the low-frequency acoustic envelopes of both nasal and affricate sounds; however it does better in matching the high-frequency envelope in nasal compared to affricate sound.

We also present the box plots to visualize statistical distribution of the SED features for nasals and affricates in Fig. 10. The box plot depicts minimum, first quartile, median, third quartile, and maximum of the data. Note that SED features deviate along frequency bands and the range of deviation is much higher for the affricates, which makes estimation of excitation for affricates harder compared to nasals.

## B. Subjective Evaluations

We performed a subjective A/B comparison test to measure the perceived quality of the proposed TM enhancement schemes. During the tests, the subjects are asked to indicate their preference for each of given A/B test pair 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. We include the AM, TM and six of the proposed enhancement configurations, and 20 comparison pairs into the A/B test. Then 30 randomly selected sentence pairs have been used in the test and the test is performed over 15 subjects. The eight conditions are numbered and the average preference scores for all comparison pairs are presented in Table IV. Note that the columns and the rows of Table IV correspond to A and B of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection.

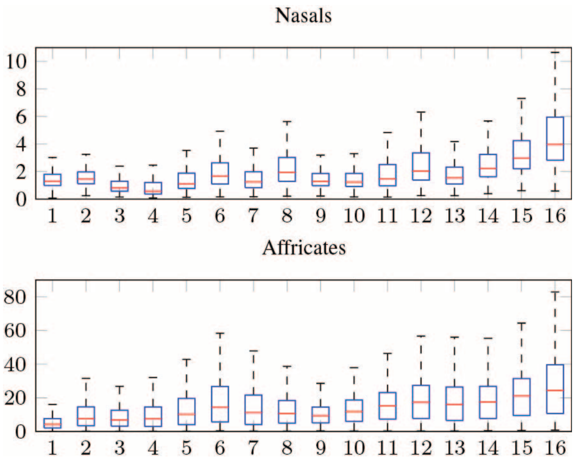

Fig. 10. Box plot of the SED features for nasals and affricates.

TABLE IV
THE AVERAGE PREFERENCE RESULTS OF THE SUBJECTIVE A/B PAIR COMPARISON TEST

| **B** | | **A** | | | | |
|---|---|---|---|---|---|---|
| # | Condition | **1** | **2** | **3** | **4** | **5** |
| **1** | $(H_T, E_T)$ | 0.00 | | | | |
| **2** | $(H_A, E_A)$ | **2.00** | 0.00 | | | |
| **3** | $(\widehat{H}_A^{H_T\|p}, \widehat{E}_A^{\mathbf{c}H_T\|p})$ | **1.75** | -1.50 | | | |
| **4** | $(\widehat{H}_A^{H_T\|p}, \widehat{E}_A^{H_T\|p})$ | **1.85** | -1.73 | -0.68 | | |
| **5** | $(\widehat{H}_A^{H_T\|p}, \widehat{E}_A^{F\|p})$ | **1.92** | -1.12 | **0.96** | **1.08** | **0.20** |
| **6** | $(\widehat{H}_A^{H_T}, \widehat{E}_A^{\mathbf{c}H_T})$ | **0.75** | -1.90 | | | -1.80 |
| **7** | $(H_T, \widehat{E}_A^{\mathbf{c}H_T})$ | **1.43** | | | | -1.70 |
| **8** | $(\widehat{H}_A^{H_T\|p}, E_T)$ | **1.24** | | | | -1.50 |

The first two conditions $(H_T, E_T)$ and $(H_A, E_A)$ are respectively the TM and AM speech. Conditions 3, 4 and 5 are the proposed phone-dependent mappings and they only differ by their excitation mappings. Condition 3 is $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{\mathbf{c}H_T|p})$, where SED features are estimated from the excitation cepstrum and LSF observations of TM. Condition 4 is $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{H_T|p})$, where SED features are estimated from the LSF observations of TM. Condition 5 is $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{F|p})$ and it has the proposed fusion mapping for the excitation as defined in (12). Furthermore, Condition 6 is the phone-independent mapping $(\widehat{H}_A^{H_T}, \widehat{E}_A^{\mathbf{c}H_T})$. Condition 7 is $(H_T, \widehat{E}_A^{\mathbf{c}H_T})$ and it uses the original TM spectral envelope with the enhanced excitation spectrum. Finally condition 8 is $(\widehat{H}_A^{H_T|p}, E_T)$ and it uses the original TM excitation with the enhanced spectral envelope.

The TM speech, condition 1, is compared to all other conditions, and it is strongly not preferred for all conditions. The AM speech, condition 2, is compared to four of the conditions excluding itself, and it is preferred for all conditions. The proposed phone-dependent fusion mapping for excitation with the best blind spectral envelope enhancement, condition 5, is preferred over all conditions except the AM speech. The phone-independent mapping, condition 6, has poor preference results as referenced to the AM speech with $-1.90$ and to the proposed phone-dependent fusion mapping with $-1.80$. Likewise the results given Table II, enhancement of excitation without any spectral envelope mapping, condition 7, has more contribution to the quality of speech than enhancement of spectral envelope without any excitation mapping, condition 8, when compared to the TM speech. Furthermore, phone-dependent mappings in conditions 3 and 4 exhibit preference scores in line with the objective scores that we present in Fig. 7. Speech samples from the listening tests are also available online [39].

## IV. Conclusion

In this paper, we developed an enhancement system for the TM recordings through source and filter separation. We extract GMM-based statistical mappings for enhancement of the spectral envelope and excitation signals over parallel recordings of AM and TM recordings. We investigate phone-dependent mappings to address the phone-dependent variability of tissue conduction with TM recordings, and we report significant performance improvements with the phone-dependent models over phone-independent models. In subjective evaluations the proposed phone-dependent enhancement, $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{F|p})$, has been preferred over the phone-independent enhancement, $(\widehat{H}_A^{H_T}, \widehat{E}_A^{\boldsymbol{\mathcal{C}}H_T})$, with preference score $-1.80$, while in the same test the AM speech has been preferred over the phone-independent enhancement with preference score $-1.90$. Furthermore, in this paper we introduce a novel excitation enhancement structure, and in subjective evaluations the proposed phone-dependent enhancement, $(\widehat{H}_A^{H_T|p}, \widehat{E}_A^{F|p})$, has been preferred over the envelope-only enhancement, $(\widehat{H}_A^{H_T|p}, E_T)$, with preference score $-1.50$. These two observations from the subjective evaluations are also synchronously supported with the objective evaluation results in terms of the PESQ, LSD and MSE metrics.

## References

[1] P. Fabre, “Un procédé électrique percutané d’inscription de l’accolement glottique au cours de la phonation: Glottographie de haute fréquence,” *Bull. de l’Acad. Nat. Méd.*, vol. 141, pp. 66–69, 1957.

[2] K. Brady, T. Quatieri, J. Campbell, W. Campbell, M. Brandstein, and C. Weinstein, “Multisensor MELPe using parameter substitution,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, vol. 1, p. I–477–80, vol.1.

[3] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, “Accurate hidden markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation,” in *Proc. IEEE Workshop Autom. Speech Recogn. Understand. (ASRU)*, Nov. 2003, pp. 73–76.

[4] G. Hugo, “Acoustic apparatus,” U.S. patent 1,521,287, 1924.

[5] G. E. Lancioni, N. N. Singh, M. F. O’Reilly, J. Sigafoos, G. Ferlisi, G. Ferrarese, V. Zullo, and D. Oliva, “A voice-sensitive microswitch for a man with amyotrophic lateral sclerosis and pervasive motor impairment,” *Disability Rehab. : Assistive Technol.*, vol. 9, no. 3, pp. 260–263, 2014.

[6] S. Roucos, V. Viswanathan, C. Henry, and R. Schwartz, “Word recognition using multisensor speech input in high ambient noise,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1986, vol. 11, pp. 737–740.

[7] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, “Combining standard and throat microphones for robust speech recognition,” *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 72–74, Mar. 2003.

[8] L. Neumeyer and M. Weinrraub, “Probabilistic optimum filtering for robust speech recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1994, pp. 417–420.

[9] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, “Air- and bone-conductive integrated microphones for robust speech detection and enhancement,” in *Proc. IEEE Workshop Autom. Speech Recogn. Understand. (ASRU)*, Nov. 2003, pp. 249–254.

[10] J. Droppo, L. Deng, and A. Acero, “Evaluation of splice on the aurora 2 and 3 tasks,” in *Proc. Int. Conf. Spoken Lang. Process.*, Sep. 2002, Int. Speech Commun. Assoc..

[11] A. Subramanya, L. Deng, Z. Liu, and Z. Zhang, “Multi-sensory speech processing: Incorporating automatically extracted hidden dynamic information,” in *Proc. Int. Conf. Multimedia Expo (ICME)*, Jul. 2005, pp. 1074–1077.

[12] J. Hershey, T. Kristjansson, and Z. Zhang, “Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition,” in *Proc. ISCA Tutorial Research Workshop Statist. Percept. Audio Process.*, Oct. 2004.

[13] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, “Unvoiced speech recognition using tissue-conductive acoustic sensor,” *EURASIP J. Adv. Signal Process.*, no. 1, pp. 56–66, 2007.

[14] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, “Multi-sensory microphones for robust speech detection, enhancement and recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, vol. 3, pp. 781–784.

[15] S. Dupont, C. Ris, and D. Bachelart, “Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise,” in *Proc. ISCA Workshop Robustness Issues Convers. Interact.*, Aug. 2004.

[16] S. C. Jou, T. Schultz, and A. Waibel, “Whispery speech recognition using adapted articulatory features,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, pp. 1009–1012.

[17] T. Dekens, W. Verhelst, F. Capman, and F. Beaugendre, “Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection,” in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2010, pp. 23–27.

[18] E. Erzin, “Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1316–1324, Sep. 2009.

[19] T. Quatieri, K. Brady, D. Messing, J. Campbell, W. Campbell, M. Brandstein, C. Weinstein, J. Tardelli, and P. Gatewood, “Exploiting nonacoustic sensors for speech encoding,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 533–544, Mar. 2006.

[20] W. Campbell, T. Quatieri, J. Campbell, and C. Weinstein, “Multimodal speaker authentication using nonacoustic sensors,” in *Proc. Workshop Multimodal User Authent.*, Santa Barbara, CA, USA, Dec. 2003, Tech. Rep..

[21] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[22] K. Kondo, T. Fujita, and K. Nakagawa, “On equalization of bone conducted speech for improved speech quality,” in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Aug. 2006, pp. 426–431.

[23] O. Makeyev, E. Sazonov, S. Schuckers, P. Lopez-Meyer, T. Baidyk, E. Melanson, and M. Neuman, “Recognition of swallowing sounds using time-frequency decomposition and limited receptive area neural classifier,” in *Applications and Innovations in Intelligent Systems XVI*, T. Allen, R. Ellis, and M. Petridis, Eds. London, U.K.: Springer, 2009, pp. 33–46.

[24] E. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. Melanson, and M. Neuman, “Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 626–633, Mar. 2010.

[25] W. Walker and D. Bhatia, "Towards automated ingestion detection: Swallow sounds," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2011, pp. 7075–7078.

[26] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, 2003.

[27] C. Yagli, M. A. T. Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a Viterbi path," *Speech Commun.*, vol. 55, pp. 111–118, Jan. 2013.

[28] G. Miet, A. Gerrits, and J. Valiere, "Low-band extension of telephone-band speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2000, pp. 1851–1854.

[29] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[30] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[31] A. Shahina and B. Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 2, pp. 1–10, Jun. 2007, Article ID 1317051.

[32] M. A. T. Turan and E. Erzin, "Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 7049–7053.

[33] M. A. T. Turan and E. Erzin, "A new statistical excitation mapping for enhancement of throat microphone recordings," in *Proc. INTERSPEECH: Annu. Conf. Int. Speech Commun. Assoc.*, 2013.

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[35] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[36] Y. Agiomyrgiannakis and Y. Stylianou, "Conditional vector quantization for speech coding," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 2, pp. 377–386, Feb. 2007.

[37] O. Salor, B. Pellom, T. Ciloglu, K. Hacioglu, and M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 349–352.

[38] ITU, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T, Tech. Rep., 2005.

[39] M. A. T. Turan and E. Erzin, "Speech samples of source and filter estimation for throat-microphone speech enhancement," [Online]. Available: http://home.ku.edu.tr/eerzin/t2a Feb. 2015

**M. A. Tuğtekin Turan** (S'11) received the B.Sc. and M.Sc. degrees, respectively, from Bilkent University, Ankara, Turkey, in 2011, and the Koç University, Istanbul, Turkey, in 2013, all in electrical engineering. He is currently pursuing the Ph.D. degree in the Electrical Engineering Department of Koç University and his research interests include articulatory phonetics, dietary monitoring, ubiquitous sensing for health applications, machine learning techniques for non-acoustic sensors and speech enhancement.

**Engin Erzin** (S'88–M'96–SM'06) received his Ph.D. degree, M.Sc. degree, and B.Sc. degree from the Bilkent University, Ankara, Turkey, in 1995, 1992 and 1990, respectively, all in electrical engineering. During 1995–1996, he was a Postdoctoral Fellow in Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he is with the Electrical & Electronics Engineering and Computer Engineering Departments of Koç University, Istanbul, Turkey. His research interests include speech signal processing, audio-visual signal processing, human–computer interaction and pattern recognition. He has served as an Associate Editor of the IEEE Transactions on Audio, Speech, Language Processing (2010–2014) and as a member in the IEEE Signal Processing Education Technical Committee (2005–2009). He was elected as the Chair of the IEEE Turkey Section in 2008–2009.