ENHANCEMENT OF THROAT MICROPHONE RECORDINGS BY LEARNING PHONE-DEPENDENT MAPPINGS OF SPEECH SPECTRA

M.A. Tuğtekin Turan and Engin Erzin

Multimedia, Vision and Graphics Laboratory
College of Engineering
Koç University, Istanbul, Turkey
mturan,eerzin@ku.edu.tr

ABSTRACT

We investigate spectral envelope mapping problem with joint analysis of throat- and acoustic-microphone recordings to enhance throat-microphone speech. A new phone-dependent GMM-based spectral envelope mapping scheme, which performs the minimum mean square error (MMSE) estimation of the acoustic-microphone spectral envelope, has been proposed. Experimental evaluations are performed to compare the proposed mapping scheme to the state-of-the-art GMM-based estimator using both objective and subjective evaluations. Objective evaluations are performed with the log-spectral distortion (LSD) and the wideband perceptual evaluation of speech quality (PESQ) metrics. Subjective evaluations are performed with the A/B pair comparison listening test. Both objective and subjective evaluations yield that the proposed phone-dependent mapping consistently improves performances over the state-of-the-art GMM estimator.

Index Terms— throat-microphone, speech enhancement, spectral envelope estimation

1. INTRODUCTION

Throat microphones capture speech signals in the form of vibrations through skin-attached piezo-electric sensors. Hence, they capture a lower bandwidth speech signal compared to acoustic-microphone recordings. Throat-microphone recordings are significantly more robust to environmental noise conditions, however they suffer from the perceived speech quality. Since throat-microphone recordings are strongly robust and highly correlated with the acoustic speech signal, they are attractive candidates for robust speech processing applications under adverse noise conditions, such as airplane, motorcycle, military field, factory or street crowd environments.

The use of non-acoustic sensors in multi-sensory speech processing has been studied for speech enhancement, robust speech modeling and improved speech recognition [1, 2, 3, 4]. In one of the early studies, throat- and acoustic-microphone speech recordings were linearly combined to enhance a noisy speech signal for improved speech recognition [5]. A device that combines a close-talk and a bone-conductive microphone is proposed by the Microsoft research group for speech detection using a moving-window histogram [6]. Direct filtering, which is based on learning mappings in a maximum likelihood framework, is investigated in [7]. Later, direct filtering is improved to deal with the environmental noise leakage into the bone sensor and with the teethclack problem [8].

Multi-sensory speech processing for noisy speech enhancement and improved noise robust speech recognition are discussed in [1, 9, 10]. 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 [11].

Graciarena et al. propose estimation of clean acoustic speech features using the probabilistic optimum filter (POF) mapping with combined throat and acoustic microphone recordings [12]. The POF mapping is a piecewise linear transformation applied to noisy feature space to estimate the clean feature space [13]. In [4], we develop a framework to define a temporal correlation model between simultaneously recorded throat- and acoustic-microphone speech. The resulting temporal correlation model is then employed to estimate acoustic features, which are spectrally richer than throat features, from throat features through linear prediction analysis. The throat microphone features and the estimated acoustic features are then used in a multimodal speech recognition system.

Non-acoustic sensors can reveal speech attributes that are lost in the noisy acoustic signal such as, low-energy consonant voice bars, nasality, and glottalized excitation. Quatieri et al. investigate methods of fusing non-acoustic low-frequency and pitch content with acoustic-microphone content for low-rate coding of speech [2].

Although throat-microphone recordings are robust to acoustic noise and reveal certain speech attributes, they often lack naturalness and intelligibility. There have been a few attempts in the literature that improve the perceived speech quality of non-acoustic sensor recordings. A neural network based mapping of the speech spectra from throat-microphone to acoustic-microphone recordings has been investigated in [14]. Note that speech spectra mapping techniques have been also studied extensively for the artificial bandwidth extension of telephone speech [15, 16]. In another study [17], the transfer characteristics of bone-conducted and acoustic-microphone speech signals are modeled as dependent sources, and an equalizer, which is trained using simultaneously recorded acoustic and bone-conducted microphone speech, has been investigated to enhance bone-conducted speech.

In this paper we target to enhance the naturalness and the intelligibility of throat-microphone speech by mapping the throat-microphone speech spectra closer to the acoustic-microphone speech spectra with a speaker- and phone-dependent probabilistic estimator, which is trained using simultaneously recorded acoustic- and throat-microphone speech. The main contribution of this paper, over

the state-of-the-art spectral mapping techniques that are investigated in [14, 15, 16, 17], is the phone-dependent spectral mapping. We observe significant improvements when the true phone-context is available to the spectral mapping. Based on this observation, we investigate phone-dependent spectral mapping in the presence of predicted phone-context. The proposed throat-microphone speech enhancement system is given in section 2. In section 3, experimental evaluations and results are given. Finally, section 4 includes the discussions and future research directions.

2. ENHANCEMENT OF THROAT-MICROPHONE SPEECH

Let us consider having two simultaneously recorded throat- and acoustic-microphone speech recordings as $s_T[n]$ and $s_A[n]$, respectively. Source-filter separation through the linear prediction filter model of speech can be defined as,

$$S_T(z) = \frac{1}{W_T(z)} R_T(z) \quad (1)$$

$$S_A(z) = \frac{1}{W_A(z)} R_A(z), \quad (2)$$

where $W_T(z)$ and $W_A(z)$ are the inverse linear prediction filters, and $R_T(z)$ and $R_A(z)$ are the source spectra for the throat- and acoustic-microphone speech, respectively. Then we can define the problem under investigation as finding a mapping from throat-microphone spectra to acoustic-microphone spectra, such as,

$$\hat{W}_A(z) = \varphi(W_T(z)|\Lambda_{TA}), \quad (3)$$

where Λ_{TA} is a general correlation model of throat- and acoustic-microphone speech, which can be extracted using a simultaneously recorded training database. Replacing the throat-microphone speech spectra with the estimated spectral envelope,

$$\hat{S}_A(z) = \frac{1}{\hat{W}_A(z)} R_T(z), \quad (4)$$

is expected to enhance the perceived quality of the throat-microphone speech.

In this study, the line spectrum frequency (LSF) feature vector representation of the linear prediction filter is used to model the spectral envelope. The throat- and acoustic-microphone spectral representations are extracted as 16th order linear prediction filters over 10 ms time frames. Let us define the elements of this representation at time frame k as column vectors $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$, respectively representing the throat-microphone spectra as an observable source $\mathcal{X}$ and acoustic-microphone spectra as a hidden source $\mathcal{Y}$.

2.1. GMM-based Mapping

The Gaussian mixture model (GMM) estimator of [18] is a soft mapping from observable source $\mathcal{X}$ to hidden source $\mathcal{Y}$ with an optimal linear transformation in the MMSE sense. It can be formulated as the MMSE estimator, which is a soft mapping (SM) from the observable source to the hidden source,

$$\hat{\boldsymbol{y}}_k^s = \sum_{l=1}^{L} p(\gamma_l|\boldsymbol{x}_k)[\boldsymbol{\mu}_{y,l} + \boldsymbol{C}_{yx,l}(\boldsymbol{C}_{xx,l})^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_{x,l})], \quad (5)$$

where γ_l is the l-th Gaussian mixture and L represents the total number of Gaussian mixtures. The vectors $\boldsymbol{\mu}_{x,l}$ and $\boldsymbol{\mu}_{y,l}$ are respectively the centroids for the l-th Gaussian for sources $\mathcal{X}$ and $\mathcal{Y}$, $\boldsymbol{C}_{xx,l}$ is the covariance matrix of source $\mathcal{X}$ in the l-th Gaussian, and $\boldsymbol{C}_{yx,l}$ is the cross-covariance matrix of sources $\mathcal{X}$ and $\mathcal{Y}$ for the l-th Gaussian mixture. The probability of the l-th Gaussian mixture given the observation $\boldsymbol{x}_k$ is defined as the normalized Gaussian pdf as,

$$p(\gamma_l|\boldsymbol{x}_k) = \frac{\mathcal{N}(\boldsymbol{x}_k; \boldsymbol{\mu}_{x,l}, \boldsymbol{C}_{xx,l})}{\sum_{m=1}^{L} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{\mu}_{x,m}, \boldsymbol{C}_{xx,m})}. \quad (6)$$

The GMM estimator can also be formulated as a hard mapping (HM) from the observable source $\mathcal{X}$ to the hidden source $\mathcal{Y}$ as,

$$\hat{\boldsymbol{y}}_k^h = p(\gamma_{l^*}|\boldsymbol{x}_k)[\boldsymbol{\mu}_{y,l^*} + \boldsymbol{C}_{yx,l^*}(\boldsymbol{C}_{xx,l^*})^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_{x,l^*})], \quad (7)$$

where γ_{l^*} represents the most likely mixture component, that is,

$$l^* = \arg\max_l p(\gamma_l|\boldsymbol{x}_k). \quad (8)$$

2.2. Phone-Dependent Mapping

Throat-microphone recordings reveal certain speech attributes, and deliver varying perceptual quality for different sound vocalizations, such as, nasals, plosives, fricatives. Hence an acoustic phone-dependent relationship between throat- and acoustic-microphone speech can be formulated to value the attributes of the throat-microphone speech. In order to explore such a relationship between throat- and acoustic-microphone speech, we first define a phone-dependent soft mapping (PDSM),

$$\hat{\boldsymbol{y}}_k^{s|c} = \frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L_n} p(\gamma_l^{c_n}|\boldsymbol{x}_k)[\boldsymbol{\mu}_{y,l}^{c_n} + \boldsymbol{C}_{yx,l}^{c_n}(\boldsymbol{C}_{xx,l}^{c_n})^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_{x,l}^{c_n})], \quad (9)$$

where N is the number of phones and each phone c_n has a separate GMM, which is defined by phone-dependent mean vectors and covariance matrices.

Furthermore, a phone-dependent hard mapping (PDHM) can be defined as,

$$\hat{\boldsymbol{y}}_k^{h|c} = \sum_{l=1}^{L^*} p(\gamma_l^{c^*}|\boldsymbol{x}_k)[\boldsymbol{\mu}_{y,l}^{c^*} + \boldsymbol{C}_{yx,l}^{c^*}(\boldsymbol{C}_{xx,l}^{c^*})^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_{x,l}^{c^*})], \quad (10)$$

where c^* is the given phone. In this study we consider three different sources for the given phone. The true phone, c^T, is defined as the true phonetic class of the phone, which is considered as the most informative upper bound for the phone-dependent model. Force alignment is used to extract the true phone source. The likely phone from the GMM, c^G, is defined as the most likely phonetic class, which can be extracted as,

$$c^G = \arg\max_{c_n} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{\mu}_{x,l}^{c_n}, \boldsymbol{C}_{xx,l}^{c_n}). \quad (11)$$

Finally, the likely phone from an HMM-based phoneme recognizer, c^M, is defined as the most likely phonetic class, which is decoded by an HMM-based phoneme recognition over the observable source $\mathcal{X}$.

3. EXPERIMENTAL EVALUATIONS

We perform experiments on a synchronous throat and acoustic microphone database which consists of 799 sentences from one male

Table 1. The Turkish METUbet phonetic alphabet with classification into 8 phonetic attributes.

Back Vowels		Stops		Fricatives	
AA	**a**nı	B	**b**al	H	**h**asta
A	l**a**f	D	**d**e**d**e	J	müj**d**e
I	**ı**sı	GG	kar**g**a	F	**f**asıl
O	s**o**ru	G	**g**enç	S	**s**e**s**
U	k**u**lak	KK	a**k**ıl	SH	a**ş**ı
Front Vowels		K	**k**edi	VV	**v**ar
E	**e**lma	P	i**p**	V	ta**v**uk
EE	der**e**	T	ütü	Z	a**z**ık
IY	s**i**m**i**t	Liquids		ZH	yo**z**
OE	**ö**rtü	LL	ku**l**	Affiricates	
UE	**ü**mit	L	**l**ey**l**ek	C	**c**am
Nasals		RR	I**r**mak	CH	se**ç**im
M	da**m**	RH	bi**r**	Glide	
NN	a**n**i	R	**r**af	Y	**y**at
N	sü**n**gü				

speaker at 16-kHz sampling rate. At the training stage, codebooks are established via varying number of Gaussian mixtures model using 720 recordings. The rest of our database are used for test stage.

The recordings are phonetically transcribed using the Turkish phonetic dictionary METUbet [19], and the phone level alignment is performed using force alignment and visual inspection. The METUbet phonetic alphabet is given in Table 1. As Salor et al. suggest in [19], the Turkish phoneme GH, soft g, has not been used for transcription and recognition, since it is used for lengthening of the previous vowel sound.

Evaluations of the spectral envelope estimation for the throat-microphone speech enhancement are performed with two distinct objective metrics, the logarithmic spectral distortion (LSD) and the perceptual evaluation of wideband speech quality (PESQ) metrics. The logarithmic spectral distortion (LSD) is a widely used metric for speech spectral envelope quality assessment. The LSD metric assesses the quality of the estimated spectral envelope with respect to the original wideband spectral envelope, and is defined as

$$d_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log\frac{|W_A(\omega)|^2}{|\hat{W}_A(\omega)|^2}\right)^2 d\omega}, \qquad (12)$$

where $W_A(\omega)$ and $\hat{W}_A(\omega)$ represent the original and estimated acoustic spectral envelopes, respectively. The ITU-T Standard PESQ [20] is employed as the second objective metric to evaluate the perceptual quality of the enhanced throat-microphone speech signal, which is constructed using the estimated spectral envelope and the excitation signal of the throat-microphone speech.

3.1. Observations on Throat-Microphone Speech Attributes

The articulation of different phones come with its distinct character in terms of resonance shaping. Although they differ in realization across individual speakers, the tongue shape and positioning in the oral cavity do not change significantly. Since, the throat-microphone captures a reliable low-frequency energy, it represents low-frequency events, such as nasals and voice bars, sufficiently well.

Table 2. The average LSD scores between throat- and acoustic-microphone spectrums for different phonetic attributes with relative occurrence frequencies in the test database.

Attribute	Freq (%)	LSD (dB)
Nasals	9.27	5.58
Stops	16.94	6.27
Liquids	9.59	7.05
Back Vowels	16.18	7.22
Front Vowels	13.93	7.65
Glide	2.36	7.81
Affiricate	2.72	9.54
Fricatives	11.10	11.81

In Table 2 we collect the average LSD scores between the acoustic and throat spectral envelopes, respectively $W_A(\omega)$ and $W_T(\omega)$, for the main phonetic attributes. The two lowest LSD scores occur for the nasals and stops, and the fricatives yield the highest LSD score. Note that nasals realized over closure of nasal cavity such as */m/* have smallest distortion, and fricatives realized over the friction of narrow-stream turbulent air such as */s/* have largest distortion due to its high-frequency energy. Clearly, the mapping of the throat-microphone speech spectra to the acoustic-microphone speech spectra is harder for the fricatives than for the nasals. That is one of the main reasons that we investigate a phone-dependent mapping for the enhancement of throat-microphone speech.

3.2. Objective Evaluations

In the test stage, the throat-microphone test recordings are separated into source $R_T(z)$ and filter $W_T(z)$ through linear prediction analysis. The estimated acoustic filter $\hat{W}_A(z)$ is extracted from the throat filter $W_T(z)$ using different mapping schemes. Then the enhanced throat-microphone recordings are synthesized using the estimated filter $\hat{W}_A(z)$ and the source $R_T(z)$.

Table 3. The average LSD and PESQ scores for different mapping schemes for enhancement of the throat-microphone recordings.

	LSD (dB)	PESQ (MOS-LQO)
PDHM-G	3.92	1.27
HM	3.80	1.29
SM	3.66	1.34
PDSM	3.65	1.36
PDHM-M	3.48	1.38
PDHM-T	3.18	1.43

Table 3 presents the average LSD scores between the estimated filter $\hat{W}_A(z)$ and the original acoustic filter $W_A(z)$, and the average PESQ scores between the enhanced throat-microphone recordings and the original acoustic-microphone recordings. Note that for

the increasing PESQ scores the LSD scores decrease in a consistent manner. The phone-dependent number of mixture components for the phone-dependent hard and soft mapping schemes are set as $L_n = 16$ for all phones. Similarly the number of mixture components for the GMM based hard and soft mapping schemes is set as $L = 256$. The worst performing mapping scheme is observed as the phone-dependent hard mapping when phone recognition is performed with the GMM classifier (PDHM-G). This is mainly due to the poor phone recognition performance of the GMM classifier, which attains 31.94% correct phone recognition in the test database. The soft mapping scheme SM has a performance improvement over the hard mapping scheme HM. The phone-dependent soft mapping (PDSM) performs close to the soft mapping (SM) scheme. The best LSD and PESQ scores are attained with the phone-dependent hard mapping when the true phone class is known (PDHM-T). The phone-dependent hard mapping with the HMM-based phone recognition (PDHM-M) attains a performance improvement and performs closest to the PDHM-T mapping scheme. The HMM-based phone recognition for the PDHM-M mapping is performed with 3-state and 256-mixture density phone level HMM recognizer, which is trained over the throat-microphone recordings of the 11 male speakers of the TAM database in [4]. Note that the test recordings in this study have been excluded from the training set of the phone level HMM recognizer. The HMM-based phone recognizer attains 62.22% correct phone recognition over the test database.

We observe significant performance improvement when the true phone class is known to the phone-dependent hard mapping (PDHM-T) scheme. Furthermore the phone-dependent hard mapping with a reliable phone recognition, in this case the PDHM-M mapping, attains the best blind estimation for the spectral envelope to enhance the throat-microphone recordings.

The throat-microphone recordings have a lower bandwidth at low-frequency bands compared to the reference acoustic-microphone recordings. Since the perceived intelligibility is poor for the throat-microphone recordings, the average PESQ scores stay at low values. In order to isolate the degradation, which is introduced by the throat source $R_T(z)$, we consider the case with the acoustic source $R_A(z)$ and throat filter $W_T(z)$ as a degraded speech signal. In this case we synthesized an enhanced speech signal using the estimated filter $\hat{W}_A(z)$ and the acoustic source $R_A(z)$. Table 4 presents the average PESQ scores for this investigation. Note that the PESQ scores are higher compared to Table 3. Furthermore the phone-dependent hard mapping PDHM-M scheme has the highest PESQ improvement.

Table 4. The average PESQ scores for different mapping schemes using acoustic residual.

	PESQ (MOS-LQO)
PDHM-G	1.66
HM	1.75
SM	1.97
PDSM	2.02
PDHM-M	2.16
PDHM-T	2.53

3.3. Subjective Evaluations

Since the reported PESQ scores stay at low values, a subjective evaluation of the proposed throat-microphone speech enhancement techniques is necessary to check whether the objective score improvements are subjectively perceivable. We performed a subjective A/B comparison test to evaluate the proposed enhancement techniques. During the test, the subjects are asked to indicate their preference for each given A/B test pair of sentences on a scale of (-2; -1; 0; 1; 2), where the scale corresponds to *strongly prefer A*, *prefer A*, *no preference*, *prefer B*, and *strongly prefer B*, respectively. The subjective A/B test includes 21 listeners, who compared 20 sentence pairs randomly chosen from our test database to evaluate 5 conditions. The acoustic-microphone speech condition is compared to all conditions with 1 pair. The throat-microphone speech condition is compared to all three enhancement schemes with 2 pairs. The GMM-based soft mapping scheme is compared to the phone-dependent hard mapping schemes PDHM-T and PDHM-M with 3 pairs. Finally, the PDHM-T scheme is compared to the PDHM-M scheme with 3 pairs.

Table 5 presents the average subjective preference results. The rows and the columns of Table 5 correspond to A and B conditions of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection. Speech samples from the subjective A/B comparison test are available for online demonstration [21].

All the three enhancement schemes yield a perceivable difference compared to the throat-microphone speech. Among the three enhancement schemes, the PDHM-T, which uses the true phone class, has the highest perceivable improvement. The proposed PDHM-M scheme has the second best perceivable improvement, which is inline with the objective evaluations.

Table 5. The average preference results of the subjective A/B pair comparison test.

A \ B	**A**	**T**	**SM**	**PDHM-T**	**PDHM-M**
Acoustic	**0.048**	-1.929	-1.929	-1.833	-1.833
Throat			**0.571**	**1.119**	**0.833**
SM				**0.492**	**0.270**
PDHM-T					-0.540

4. CONCLUSIONS

We introduce a new phone-dependent GMM-based spectral envelope mapping scheme to enhance throat-microphone speech using joint analysis of throat- and acoustic-microphone recordings. The proposed spectral mapping scheme performs the minimum mean square error (MMSE) estimation of the acoustic-microphone spectral envelope within the phone class neighborhoods. Objective and subjective experimental evaluations indicate that the phone-dependent spectral mapping yields perceivable improvements over the state-of-the-art phone independent mapping schemes. Overall, the proposed phone-dependent spectral mapping PDHM-M introduces a significant intelligibility improvement over the throat-microphone speech. However, there is still a big room to further improve the perceive quality by modeling the source excitation signal of the throat-microphone recordings.

5. REFERENCES

[1] Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, and Alex Acero, "Speech modeling with magnitude-normalized complex spectra and its application to multisensory speech enhancement," in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 1157–1160.

[2] T.F. Quatieri, K. Brady, D. Messing, J.P. Campbell, W.M. Campbell, M.S. Brandstein, C.J. Weinstein, J.D. Tardelli, and P.D. Gatewood, "Exploiting Nonacoustic Sensors for Speech Encoding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 533–544, Mar. 2006.

[3] Szu-Chen Jou, Tanja Schultz, and Alex Waibel, "Whispery Speech Recognition using Adapted Articulatory Features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. I, pp. 1009–1012.

[4] Engin Erzin, "Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1316–1324, Sept. 2009.

[5] S. Roucos, V. Viswanathan, C. Henry, and R. Schwartz, "Word recognition using multisensor speech input in high ambient noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1986, vol. 11, pp. 737–740.

[6] Yanli Zheng, Zicheng Liu, Zhengyou Zhang, Mike Sinclair, Jasha Droppo, Li Deng, Alex Acero, and Xuedong Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 249–254.

[7] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct Filtering for Air- and Bone-Conductive Microphones," in *IEEE 6th Workshop on Multimedia Signal Processing*, 2004, pp. 363–366.

[8] Zicheng Liu, Amar Subramanya, Zhengyou Zhang, Jasha Droppo, and Alex Acero, "Leakage Model and Teeth Clack Removal for Air- and Bone-Conductive Integrated Microphones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[9] Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, Jasha Droppo, and Alex Acero, "A Graphical Model for Multi-Sensory Speech Processing in Air-and-Bone Conductive Microphones," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2005.

[10] A. Subramanya, Li Deng, Zicheng Liu, and Zhengyou Zhang, "Multi-Sensory Speech Processing: Incorporating Automatically Extracted Hidden Dynamic Information," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, 2005, pp. 1074–1077.

[11] Stephane Dupont, Christophe Ris, and Damien Bachelart, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise .," in *Proc. of Robust 2004 (Workshop on Robustness Issues in Conversational Interaction)*, 2004.

[12] Martin Graciarena, Horacio Franco, Kemal Sonmez, and Harry Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003.

[13] Leonard Neumeyer and Mitchel Weinrraub, "Probabilistic optimum filtering for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. 417–420.

[14] A. Shahina and B. Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

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

[16] Can Yagli, M.A. Tugtekin Turan, and Engin Erzin, "Artificial bandwidth extension of spectral envelope along a Viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, Jan. 2013.

[17] Kazuhiro Kondo, Tomoe Fujita, and Kiyoshi Nakagawa, "On Equalization of Bone Conducted Speech for Improved Speech Quality," in *IEEE International Symposium on Signal Processing and Information Technology*, Aug. 2006, pp. 426–431.

[18] Yannis Agiomyrgiannakis and Yannis Stylianou, "Conditional Vector Quantization for Speech Coding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 377–386, Feb. 2007.

[19] Ozgul Salor, Bryan Pellom, Tolga Ciloglu, Kadri Hacioglu, and Mübeccel Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language," in *International Conference on Spoken Language Processing (ICSLP02)*, 2002, pp. 349–352.

[20] ITU-T Recommendation P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Tech. Rep., ITU, 2005.

[21] M.A.T. Turan and E. Erzin, "Speech samples for the throat-microphone speech enhancement schemes. http://home.ku.edu.tr/˜eerzin/t2a-icassp13," Nov. 2012.