An Analytic Modeling Approach to Enhancing Throat Microphone Speech Commands for Keyword Spotting

Jun Cai[1], *Stefano Marini*[1], *Pierre Malarme*[1], *Francis Grenez*[1], *Jean Schoentgen*[1, 2]

[1] Faculté des Sciences Appliquées, Université Libre de Bruxelles, Belgium
[2] National Fund for Scientific Research, Belgium
{Jun.Cai, smarini, Pierre.Malarme, fgrenez, jschoent}@ulb.ac.be

Abstract

This research was carried out on enhancing throat microphone speech for noise-robust speech keyword spotting. The enhancement was performed by mapping the log-energy in the Mel-frequency bands of throat microphone speech to those of the corresponding close-talk microphone speech. An analytic equation detection system, Eureqa, which can infer nonlinear relations directly from observed data, was used to identify the enhancement models. Speech recognition experiments with the enhanced throat microphone speech keywords indicate that the analytic enhancement models performed well in terms of recognition accuracy. Unvoiced consonants, however, could not be enhanced well enough, mostly because they were not effectively recorded by the throat microphone.

Index Terms: speech enhancement, throat microphone, speech keyword spotting, Eureqa system

10.21437/Interspeech.2010-329

1. Introduction

As an attempt to facilitate medical services with speech recognition techniques, this research aimed at developing a keyword spotting system as a natural user interface in operating rooms. The system was intended to help the surgeons to set up surgical devices, to access media, to control the surgical data recording system, and to tag medical data or define operating steps. During the operation, the system picks certain speech commands out from the user's various speeches. Normal air-conductive (AC, hereafter) microphones cannot be used effectively for this system because usually they would be exposed to loud and sharp ambient noises during surgery. Furthermore, like other devices in the operating room, the microphone in this system needs to be sterilized frequently. Sterilization of the AC microphone is indeed a complex task. To avoid these problems, body-conductive (hereafter BC) microphone could be used instead to acquire the speech signal.

BC microphones can pick up speaker's vocal fold vibrations (phonation) clearly without distortion through the conduction of the skin or bone around the neck. They are designed to be environmental air vibration proof. Therefore, the background noises can be effectively eliminated in the speech picked up by BC microphones. Different types of BC microphones have been invented, such as the throat microphone [1, 2], the bone-conductive microphone [3, 4], and the soft-tissue-conductive NAM microphone [5, 6]. On the market, the products of throat microphones have already been used to improve the speech intelligibility for speech communication in very loud backgrounds [7].

Yet, the sound a BC microphone acquires is different from the air-transmitted speech. Speech variations produced by the articulators such as tongue, jaw, and lips in the frontal section of vocal tract are muffled and weak in the BC microphone speech. The sound of consonants, especially that of the unvoiced fricatives, cannot be heard clearly from the BC speech signal. Therefore, in comparison to the AC microphone speech, the BC microphone speech is of low quality in terms of speech naturalness. This weakness of the BC microphone prevents it from being widely accepted in speech communication and recognition.

To deal with this problem, several speech enhancement techniques have been proposed to convert the BC speech to clear AC speech. T. Toda, et al proposed a set of statistical sound conversion approaches to enhance various types of BC speech acquired with the NAM microphone [6, 8]. The conversions were based on several GMMs which were trained to map the feature vectors of NAM-captured speech to those of highly audible voiced speech or whisper. The enhanced speech was then synthesized by using the speech vocoder STRAIGHT with the converted feature values. This speech enhancement technology can yield significant improvements in the naturalness of the BC voiced speech. A. Shahina, et al [9] proposed a speaker-dependent mapping between the spectral features of the BC speech and the AC speech by using a forward-propagation neural network. The spectra of the reconstructed speech showed that the high frequency components that were previously of low amplitude in the BC speech were well emphasized by the mapping. As a result, the perceptual quality of the BC speech was improved.

In this research, the throat microphone was used because of its ease of application and its availability on the market. Two simultaneous speech corpora, one of speech commands and the other of continuous speech sentences, have been recorded via the throat microphone and a close-talk microphone synchronously. The simultaneous recordings were windowed to form frame sequences, and the log-energy at the output of each channel of the filterbank for computing HTK's [10] MFCC was extracted frame by frame as the speech feature. Non-linear analytic models for mapping the features of throat microphone speech to those of the close-talk microphone speech were discovered by using the automatic equation detection system Eureqa [11]. The discovered models were used to enhance the throat microphone speech. To evaluate the quality of the enhancement, both speech commands and continuous speech sentences were recognized by the HTK speech recognition system.

The simultaneous corpora are described in Sec. 2. A description of a spectrogram of throat microphone speech is also conducted in this section. In Sec. 3, the feature selection scheme is presented. Sec. 4 describes the enhancement model identification by using the Eureqa system. In Sec. 5, the experiments and the results are presented. It is concluded in Sec. 6 that the proposed technique validly enhances the throat microphone speech commands for keyword spotting.

2. The simultaneous corpora of BC/AC speech

Three different throat microphones – Stryker Lite from Clearer Communications, XTM880D4 from IXRadio, and the throat microphone produced by VoiceTouch Co. – have been tested and compared in terms of the self-noise level, ambient noise sensitivity, and speech intelligibility. Among them, the one made by VoiceTouch was selected to be used in the system because it has the best comprehensive performance.

A one-speaker simultaneous corpus which consists of both the throat microphone speech commands and the corresponding close-talk microphone speech commands has been recorded in a soundproof booth. During the utterances, the speech signals acquired by both the throat microphone and the close-talk microphone were recorded synchronously by using a MOTU UltraLite-mk3 audio interface which has two input channels. The speech command set consisted of 17 commands, including "Operator", "Previous Step", "Next Step", and "Option One" to "Option Nine". A command list which contained 127 commands was uttered by the speaker and recorded. To include more phonetic variations in the corpus, 100 short sentences comprised of words out of the command set were also recorded.

In addition, a simultaneous continuous speech corpus was recorded as well, to evaluate the effectiveness and performance of the speech enhancement models. The 330 sentences in the Nov.'92 ARPA WSJ test set were read once by the same speaker and recorded. Since the entire recording process took a long time, it was divided into 8 epochs. Over all epochs, effort was made to maintain the consistency of the recording conditions, especially the consistency of the throat microphone contact point on the neck.

The spectrograms of the simultaneous recordings of

"About half these managers are in the U. S."

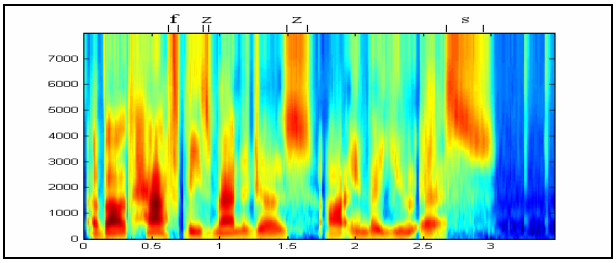
are depicted in Figure 1(a) and 1(b), respectively. Comparison of them shows that the throat microphone has a strong self-noise which frequency ranges from 3*kHz* to 8*kHz*. In the throat microphone signal, high frequency formants are very weak. The formants above 5.6*kHz*, if they do exist in the signal, are so weak that they are totally masked by the self-noise. For example, the high frequency content of the unvoiced consonant [f] almost totally disappears in Figure 1(b). This observation explains why the [f] almost cannot be heard in the throat microphone speech while it is clear in the close-talk microphone speech. In Figure 1(b), it can be also observed that the throat microphone speech is accompanied by an extra noise in the frequency range [1.3*kHz*, 1.9*kHz*]. At present, we do not know its origin, though we know that it is not an ambient noise.

In Figure 1(a), there are two bright strips at the end of the spectrogram, owing to air-borne noises. They have no counterparts in Figure 1(b) because of the ambient noise robustness of the throat microphone.

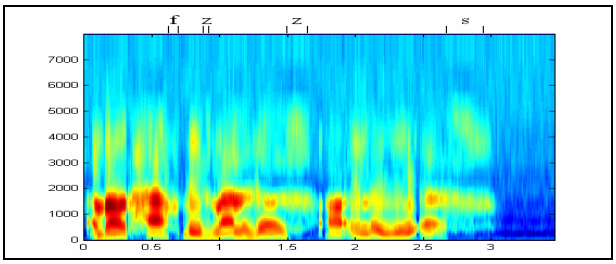
3. Speech feature selection

Within HTK system, speech signals are encoded into MFCC coefficients. Given a speech signal $x[n]$, the DFT of the signal, $X_a[k]$, must be firstly computed. A Mel-scale filterbank is then used to filter the spectrum of the signal. The log-energy at the output of each channel of the filterbank is computed as [12]:

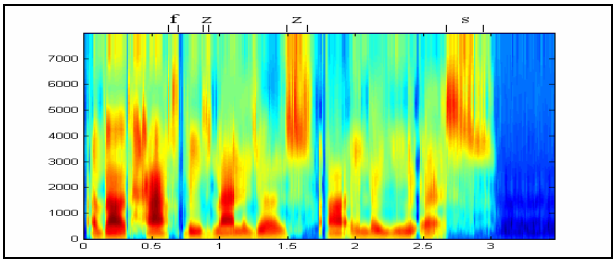
$$S[m] = \ln\left[\sum_{k=0}^{N-1} \left| X_a[k] \right|^2 H_m[k]\right], \quad 0 < m \le M \qquad (1)$$



(a) for close-talk microphone speech



(b) for throat microphone speech



(c) for enhanced speech

Figure 1: *The Spectrograms of the speech sentence "About half these managers are in the U. S."*

Here, M represents the number of channels in the filterbank; N is the number of signal samples. $H_m[k]$ is the transfer function of the m*th* channel of the filterbank. The MFCCs are then defined as the discrete cosine transform of the M filter outputs:

$$c[n] = \sum_{m=1}^{M} S[m] \cos(\pi n(m-1/2)/M), \quad 0 \le n < M \qquad (2)$$

In HTK, typically only the first 13 MFCCs are used, while M is set to 26.

Either the MFCCs $c[n]$ ($0 \le n < 13$) or the log-energy spectra $S[m]$ ($0 < m \le M$) can be used as the features for throat microphone speech enhancement dedicated to speech recognition. The MFCCs were originally selected as the features, but the enhancement models were too complex to identify accurately because each MFCC is a mixture of the outputs of all the channels. Hence, $S[m]$ ($0 < m \le M$) were selected as the features to perform speech enhancement. Since the throat microphone and the close-talk microphone are memoryless systems, it is plausible to assume that the feature mapping for the enhancement is static. Therefore, no dynamic feature such as the first- or second-order time derivatives were included in the enhancement. Only the $S[m]$'s of the throat microphone speech were mapped to those of the close-talk microphone speech. After the mapping, the MFCCs of the enhanced speech can be computed directly by using Eq. (2) with the converted feature vectors. Because the $S[m]$'s are computed after spectral filtering when deriving MFCCs, the extra computational overhead of the speech enhancement is the mapping of the features only. No other conversion is needed for obtaining MFCCs. As a consequence, the real-time

performance of the keyword spotting system is not expected to deteriorate significantly with speech enhancement.

4. Enhancement model identification

Denote the Mel-scale log-energy spectra of the throat microphone speech and those of the close-talk microphone speech as $S_t[m]$ and $S_c[m]$ ($0 < m \le M$), respectively. The enhancement models attempt to map $S_t[m]$ ($0 < m \le M$) to $S_c[m]$ ($0 < m \le M$). Figure 1 suggests that in the frequency range below 5.0*kHz*, one may hypothesize the $S_t[m]$'s are roughly weakened versions of $S_c[m]$'s. Therefore, in this band the value of a specific $S_c[m]$ can be estimated mainly from the value of the relevant $S_t[m]$. Other $S_t[m]$'s within and out of this band can also be involved in the estimation of the specific $S_c[m]$ to increase the modeling accuracy. For estimating a specific $S_c[m]$ in the frequency range above 5.0*kHz*, however, the value of the corresponding $S_t[m]$ is almost useless because not enough information is available. In this situation, we have to use the low frequency $S_t[m]$'s to estimate the high frequency $S_c[m]$.

There is no a priori knowledge about which $S_t[m]$'s may contribute to the estimation of a particular $S_c[m]$. We let the model identification algorithm determine which $S_t[k]$'s ($0 < k \le 26$) are involved in modeling each $S_c[m]$. Before model identification, all 26 $S_t[m]$'s were accepted as input variables of the model, i.e.,

$$S_c[m] = \boldsymbol{f}_m(S_t[1], S_t[1], \Lambda\ , S_t[26]), \quad 0 < m \le 26 \qquad (3)$$

Here, $\boldsymbol{f}_m(\cdot)$ is the m*th* enhancement model.

The Eureqa system has been used to identify the nonlinear enhancement models. This system can discover analytic laws that underlie physical phenomena directly from experimentally captured data by using a computational search, without any a priori knowledge about the phenomena [11]. The discovery of the analytic equation starts by searching within a dataset of observed values that seem having an intrinsic relationship to each other, then proposing a series of simple equations to describe the relation. These initial equations invariably fail, but some are slightly less wrong than others. The best ones are selected as building blocks. They are slightly changed and combined together to form more complex analytic expressions which are again tested against the data. Like genetic algorithms, Eureqa repeats the modification/combination and test cycle over and over, until it finds analytic equations that precisely describe the relationship in the dataset. For identifying the enhancement models expressed in Eq. (3), the Eureqa system not only discovers the non-linear analytic equations of the models, but also determines the input variables among $S_t[k]$'s ($0 < k \le 26$) for each model.

5. Experimental results

5.1. Continuous speech recognition

24 sentences from all the 8 epochs of the simultaneous continuous speech corpus were used as the dataset for identifying the enhancement models. The Eureqa system ran on a cluster of PCs with 26 CPU cores in total. For modeling each $S_c[m]$ ($0 < m \le M$), the search stopped after 2×10^{12} formula had been evaluated. The source code of the HTK function FBank2MFCC( ) in HSigP.c was modified to implement the enhancement of the throat microphone speech.

Speech recognition experiments using HTK on the simultaneous continuous speech corpus and the enhanced speech have been carried out to evaluate the performance of the enhancement models. The 32-Gaussian, 8000-tired-state acoustic models trained by K. Vertanen [13] on WSJ+TIMIT corpora and the WSJ standard 5K non-verbalized closed bigram language model were adopted in the experiments. The recognition accuracy is presented in Table 1.

Table 1. *Recognition accuracy for different types of continuous speech sentences.*

Type of Speech	Word Correctness (%)
WSJ Baseline	94.78
Close-talk Mic	78.49
Throat Mic	12.60
Enhanced Throat Mic	29.86

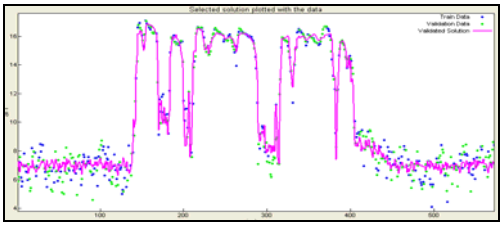
The recognition results of the sentence in Figure 1 are listed in Table 2.

Table 2. *Recognition results for the speech sentence "About half these managers are in the U. S.".*

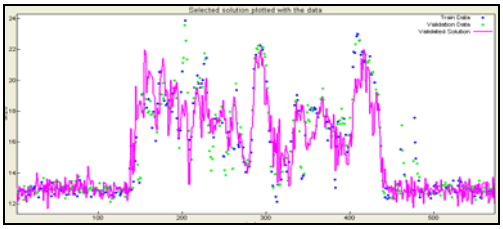
Type of Speech	Recognition result
Close-talk Mic	About how these men and is fighting the U. S.
Throat Mic	About the United Auto he added.
Enhanced Throat Mic	About a mandate is by the E. S.

Table 1 and Table 2 show that the enhancement models can improve the quality of the throat microphone speech to some extent, though the improvement is not satisfactory in the framework of continuous speech recognition.

Further analysis indicates that the formants of voiced utterances can be recovered perfectly from the throat microphone speech by the enhancement models. For unvoiced speech, though the high frequency content can be partly recovered, the enhancement does not perform well enough. Figure 1(c), which depicts the spectrogram of the enhanced version of the throat microphone speech in Figure 1(b), gives an illustration of the enhancement for both voiced and unvoiced utterances. It can be seen that the self-noise of the throat microphone and the band noise in [1.3*kHz*, 1.9*kHz*] have been well suppressed.



(a) *for the model of* $S_c[1]$



(b) *for the model of* $S_c[26]$

Figure 2: *Illustrations of model agreement*

The agreement for the models of $S_c[1]$ and $S_c[26]$ on the same speech sentence are shown in Figure 2. In the figures, the points represent the expected values, while the curves represent the values evaluated by the enhancement models. It can be observed that for the low-frequency $S_c[1]$, the

converted values match the values of the close-talk microphone speech well. But for the high-frequency $S_c[26]$, it is hard to discover models which are highly suitable.

5.2. Speech command recognition

For each keyword in the speech command simultaneous corpus, three utterances were used to extract ($S_t[m]$, $S_c[m]$) pairs. For building the enhancement models, these data were combined with the ($S_t[m]$, $S_c[m]$) pairs extracted from 20 short sentences of the 100-sentence dataset. The identified models are non-linear functions, such as:

$$S_c[1] = 0.84\sqrt{S_t^2[1] + 5.18S_t[6] + S_t[12]} - 1.48\,\mathbf{sin}(0.34S_t[2]),$$

$$S_c[2] = 0.63 + 0.89S_t[2] + 1.40\,\mathbf{log}(S_t[7]) - 15.02/S_t[12],$$

$$S_c[3] = S_t[3] + S_t[7]\sqrt{0.12S_t[7]}/S_t[3], \quad (4)$$

… … …

$$S_c[25] = -\sqrt{S_t[1]} - \mathbf{cos}(0.65S_t[12]) + 3.83S_t[12]/S_t[1] + S_t[17] + \mathbf{sin}(S_t[17]),$$

$$S_c[26] = 0.87S_t[17] - (34.64 + 6.31\,\mathbf{cos}(0.90S_t[17]) + 1.87S_t[17]\,\mathbf{cos}(0.41S_t[12]))/S_t[1].$$

The same acoustic models as in 5.1 were used to carry out recognition experiments on the simultaneous command corpus and the enhanced speech commands. The language model was a word-loop grammar of all the commands. The word accuracy of recognition is presented in Table 3. It can be seen that the word accuracy was significantly improved from 16.50% to 70.87% by enhancement.

However, the word correctness obtained by acoustic model adaptation using MLLR on the 100 short throat microphone speech sentences in the simultaneous speech command corpus was 88.20%, much higher than that obtained here by enhancement. This suggests that for the purpose of recognizing throat microphone speech, the model adaptation technique may provide a better solution.

Table 3. *Recognition accuracy for different records of speech commands.*

Type of Speech	Word Correctness (%)
Close-talk Mic	92.72
Throat Mic	16.50
Enhanced Throat Mic	70.87

Another observation was that the enhancement models being built on different corpora were different. For example, in the experiment described in 5.1, the model of $S_c[3]$ is

$$S_c[3] = 4.88 + S_t[3]\,\mathbf{hill\ 2}(0.05S_t[2] + 0.11\mathbf{cos}(S_t[2]) + 0.08\,S_t[7] - 0.14), \quad (5)$$

which differs from the model of $S_c[3]$ in Eq. (4). A possible explanation is that the mappings are similar even though their model components differ. Another reason may be that the recording conditions of the throat microphone changed slightly over different recording epochs. Using the enhancement models discovered in 5.1, the recognition of the enhanced speech commands achieved a word correctness of 68.45%, close to the word correctness using the models trained on the speech commands directly. This confirms that enhancement models discovered on one corpus can be used to enhance any speech signals recorded by the same throat microphone, provided that the recording conditions are consistent.

6. Conclusions

With the speech features being defined as the log-energies at the output of each channel of the Mel-scale filterbank for computing MFCCs, non-linear analytic models can be found by using the Eureqa system to enhance throat microphone speech. Voiced speech can be validly enhanced, while the enhancement for unvoiced speech is not good enough. This method is useful for throat microphone speech enhancement for speech recognition, though further research is needed to improve the enhancement performance.

7. Acknowledgements

This work was supported by the "Région Wallonne", Belgium, in the framework of the "WALEO II" program. The authors would like to thank Prof. Tomoki Toda of Nara Institute of Sci. and Tech., Japan for his kind help and discussion. Gratitude is also due to Christophe Mertens, Rudy Ercek and Geoffrey Vanbienne at ULB for their supports and helps.

8. References

[1] Graciarena, M., Franco, H. and Sonmez, K., et al, "Combining Standard and Throat Microphones for Robust Speech Recognition", IEEE SIGNAL PROCESSING LETTERS, 10(3): 72-74, 2003.

[2] Jou, S.-C. Schultz, T. and Waibel, A., "Adaptation for Soft Whisper Recognition Using a Throat Microphone", in Proc. INTERSPEECH 2004, Jeju Island, Korea, Oct 2004.

[3] Zheng, Y., Liu, Z. and Zhang, Z., et al, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", in Proc. ASRU 2003: 249-254, 2003.

[4] Tsuge, S., Osanai, T. and Makinae, H., et al, "Combination Method of Bone-Conduction Speech and Air-Conduction Speech for Speaker Recognition", in Proc. INTERSPEECH 2008: 1929-1932, Brisbane, Australia, Sept 2008.

[5] Nakajima, Y., Kashioka, H. and Shikano, K., et al, "Non-audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin", in Proc. ICASSP 2003: V-708 - V-711, Hong Kong, China, Apr 2003.

[6] T. Toda, K. Nakamura, T. Nagai, et al, "Technologies for Processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone", In Proc. INTERSPEECH 2009: 632-635, Brighton, UK, Sept 2009.

[7] Noe, R., "You took the words right out of my thorax: a look at throat mics", Core77 Design Magzine & Resource. Online: http://www.core77.com/blog/object_culture/you_took_the_words_right_out_of_my_thorax_a_look_at_throat_mics_13008.asp, accessed on 26 Mar 2010.

[8] Toda, T., Nakamura, K. and Sekimoto, H., et al, "Voice Conversion for Various Types of Body Transmitted Speech", In Proc. ICASSP 2009: 3601-3604, Taipei, Taiwan, Apr 2009.

[9] Shahina, A. and Yegnanarayana, B, "Mapping Speech Spectra from ThroatMicrophone to Close-Speaking Microphone: A Neural Network Approach", EURASIP Journal on Advances in Signal Processing, Vol. 2007, Article ID 87219, 2007.

[10] Young, S., Evermann, G., and Gales, M., et al, "The HTK Book", Online: http://htk.eng.cam.ac.uk/docs/docs.shtml, accessed on 15 Apr 2010.

[11] Schmidt, M. and Lipson, H., "Distilling Free-Form Natural Laws from Experimental Data", SCIENCE, 324: 81-853, Apr 2009.

[12] Huang, X., Acero, A. and Hon, H.-W., "Spoken Language Processing: A Guide to Theory, Algorithm and System Development", Prentice Hall PTR, New Jersey, 2001.

[13] Vertanen, K., "HTK Acoustic Models", Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed on 21 Mar 2010.