INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Real Time detection of soft voice for speech enhancement

Hector A. Cordourier Maruri[1], *Georg Stemmer*[2], *Sinem Aslan*[3], *Tobias Bocklet*[4], *Himanshu Bhalla*[5]

[1]Intel Corp., Mexico [2]Intel Corp., Germany
[3]Intel Corp., USA
[4]Nuremberg Tech, Germany
[5]Intel Corp., India

hector.a.cordourier.maruri@intel.com, georg.stemmer@intel.com, sinem.aslan@intel.com, tobias.bocklet@th-nuernberg.de, himanshu.bhalla@intel.com

Abstract

People in remote meetings in open spaces might choose to speak with a restrained voice due to concerns around privacy or disturbing others. Research shows that persons prefer to use soft voice (voice with lower amplitude and pitch, but with harmonic tones in its spectrum) over whispered voice (voice with the lowest amplitude, and no harmonics at all) to avoid being overheard during such calls. We present a lightweight classifier based in a simple feed-forward neural network, which uses normalized Log-Mel spectrum of voice captured by a headset as input, and can detect if the person is using soft voice. This allows to enhance soft voice with more precision and responsiveness than regular amplitude compensation ("auto-gain") systems.

In this show and tell, we present a real-time demo of the voice classifier. Viewers will see our algorithm detect in real-time soft voice vs other voice types, in a regular PC, with voice captured with a headset.

Index Terms: Para-linguistics, voice intelligibility, voice pleasantness, voice quality, remote communications, voice classification

1. Introduction and Motivation

During virtual meetings or calls in public spaces (e.g., airports, cafes, open offices, etc.), lack of privacy and the concern of disturbing others become important issues to address [1, 2], which negatively impact speakers who need to adjust their voices to preserve privacy or not to disturb others, and also affect other meeting participants who could struggle to understand the adjusted voice on the other side of the line. When people try to conceal their voice, they can change their voicing in some common ways. For example, the phonation pipeline in the vocal tract can be changed with a range that goes from dampening phonation (i.e., speaking softly), to canceling phonation (i.e., whispering) [3]. Voice production based on vocal effort can be categorized in five ascending modes: whispered, soft, neutral, loud, and shouted [4, 5, 6]. In a former study [7], we concluded that persons trying to conceal their voice from potential eavesdroppers preferred using soft voice mode, which is voice with much lower amplitude and pitch than regular voice, but that still retains harmonic tones in its spectrum. The other less favored alternative was whispered voice mode, which has the lowest amplitude, and no harmonics at all. Soft voice proved to be similarly effective in concealment on public places, and also was reported as more comfortable and less exhausting than whispered voice.

However, such study also showed that concealment behaviours made the voice less clear or more unpleasant for the persons in the other side of the line. This mainly because softer voices have significant lower amplitude, which by itself reduces intelligibility [8]. As a first step to process soft voice samples to enhance intelligibility, it is useful to be able to detect the actual presence of soft voice during a conversation. This detection routine needs to be lightweight and responsive enough to be practical during a live conversation without adding significant latency or overhead.

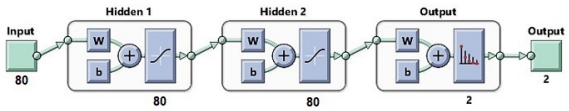

Figure 1: *Shallow neural net classifier architecture.*

We propose a real-time voice classifier based in a simple feed-forward neural network, which uses the normalized logarithmic Mel spectrum of voice captured by a headset as input, and as output determines if the person is currently using soft voice. This classifier will enable quick voice enhancements techniques, which can go from simple noise reduction and gain compensation, to sophisticated voice transformation routines.

2. Proposed classifier solution

As voice classifier, we propose a simple fully connected, shallow neural net of two layers, which receives a normalized logarithmic Mel spectrum of 80 coefficients as input, and produces two possible classes as output (soft voice / no soft voice), as can be seen in Figure 1. This configuration consists of 13k programmable parameters, which makes it relatively lightweight for most processing platforms.

To train the net, 230 minutes of audio (33.3% of soft, 33.3% of regular and 33.3% of whispered voice) were used, 80% (184 min.) of it for training and 20% (46 min) for testing. All audio was recorded with headsets in actual virtual conference situations, at 16kHz sample frequency.

The test performance of the classifier can be seen in the confusion matrix in Figure 2, for a total of 14,361 audio frames of 0.26 ms each (4096 samples), randomly selected from the testing data.

3. Real-time demo

3.1. Construction

The routine captures 16kHz sample frequency audio from a headset and obtains the normalized logarithmic 80 coefficient Mel spectra of subsequent time windows of 4096 samples, with 87.5% overlap. Then, the consecutive spectrum vectors are fed to the trained net classifier, which produces an output of proba-

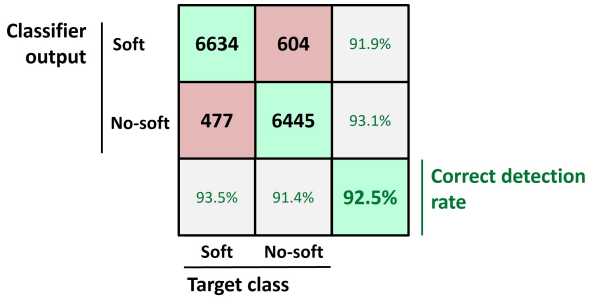

Figure 2: *Confusion matrix of the tests classification results of the trained net.*

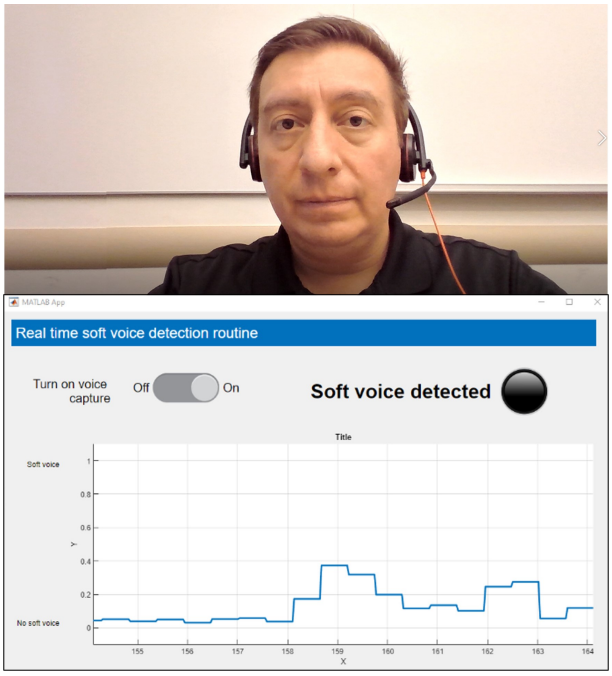

Figure 3: *Image of the graphical user interface of the show and tell demo of the real-time soft voice detection routine.*

bility of soft voice detection. The probability of past and current time windows is smoothed with a 5-point moving average, and then, if the value reaches above 0.5, the audio frame is considered positive soft voice detection.

The demonstration routine was implemented in MATLAB app designer, using the Audio Toolbox to access the audio input / output streams in a regular Windows based PC, in which the voice is captured with a Plantronics Blackwire 3220 USB wired headset, equipped with a boom mic. The implementation GUI shows a plot with the output current soft voice detection probability after the moving average, with an indicator of a positive soft voice detection. A capture of the GUI, can be seen in Figure 3.

4. Conclusions

The proposed technique is able to recognize the presence of soft voice with 92.5% success rate among regular and whispered voice. This means the proposed algorithm, while very simple in itself, is not just amplitude or power driven, but it analyzes the spectral characteristics to correctly detect soft voice frames.

The real time demo shows how the proposed routine is lightweight and also robust enough to detect the presence of soft voice in real time during ongoing speech. This classification enables future enhancement techniques for the soft voice sections of a conversation, which can go from simple gain compensations, to voice transformation routines.

5. References

[1] K. Jensen and E. Arens, "Acoustical quality in office workstations, as assessed by occupant surveys," *Proceedings, Indoor Air 2005*, 01 2005.

[2] D. A. Ilter, E. Ergen, and I. Tekce, "Acoustical comfort in office buildings," in *7th Annual International Conference -ACE 2019 Architecture and Civil Engineering At Singapore*, 05 2019.

[3] Z. Zhang, "Mechanics of human voice production and control," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2614–2635, 2016. [Online]. Available: https://doi.org/10.1121/1.4964509

[4] H. A. Patil, A. Neustein, and M. Kulshreshtha, *Signal and Acoustic Modeling for Speech and Communication Disorder*, ser. Speech Technology and Text Mining in Medicine and Health Care. De Gruyter, 2018. [Online]. Available: https://books.google.com/books?id=4-WTDwAAQBAJ

[5] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000. [Online]. Available: https://doi.org/10.1121/1.429414

[6] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763931200009X

[7] H. Cordourier-Maruri, S. Aslan, G. Stemmer, N. Alyuz, and L. Nachman, "Analysis of contextual voice changes in remote meetings," 08 2021, pp. 2521–2525.

[8] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12061