Separating known competing voices for people with hearing loss

Niels H. Pontoppidan[1], *Marianna Vatti*[1], *Rikke Rossing*[1], *Tom Barker*[2], *Tuomas Virtanen*[2]

[1] Eriksholm Research Centre, Oticon A/S, Denmark
[2] Technical University of Tampere, Finland

[1]{**npon**,mvat,riro}@eriksholm.com, [2]{thomas.barker,tuomas.virtanen}@tut.fi

Index Terms: speech recognition, competing voices, known voices separation, non-negative matrix factorization, source separation, hearing loss.

1. Introduction

In noisy situations with competing voices, people often find it easier to hear out and recognize what familiar voices say. However, people with hearing loss report difficulties in utilizing this ability in situations with competing voices. This research investigates if hearing devices can learn and utilize voice characteristics to separate voices, and furthermore, if presenting the separated voices with increased spatial separation enhances their ability to separate the competing voices.

In this study, 13 people with moderate sloping hearing loss tested a known voices separation algorithm to measure benefit from the known voices separation algorithm in a competing voices situation.

2. Theory

People with hearing loss perform significantly worse in complex listening situations as the competing voices scenario. Reduced sensitivity to weak sounds, reduced frequency selectivity, and inability to utilize temporal cues all contribute to this problem. This research explores the possibility to perform the separation on behalf of the person with hearing loss and if the hearing devices can transform the listening situation to a simpler situation with bigger segregation cues.

The known voices separation algorithm [1] investigated here uses Non-negative Matrix Factorization to construct models for the individual voices. The individual voice models follows the assumption that each voice exhibit repeatable patterns of progressing from one configuration of the voice production units to another. Thus, the characteristics of each voice is how the spectrum changes over time as seen in Figure 1 below.

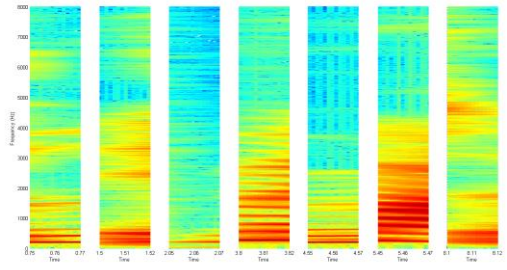

Figure 1: An example of seven elements voice model. Notice that each element would be short in practice and that at least hundred elements is required to model a voice.

Consequently, the input to the separation follows the tradition described in [2]; starting with segmenting the speech signal into overlapping timeframes, each transformed into the frequency domain, and finally grouped together such that each element is two (or more) consecutive spectra. The last step is the important step, as the consecutive spectra enables the learning of how each voice progresses from one spectrum to another, and thus the assumption that the voice can be modelled as such transitions.

Learning the individual voices can include a reduction of the number of elements in the model. While, this study does not cover this in detail, it does differ from earlier work [2] by only using 20 seconds of clean voices to build the models. Another important aspect is the delay that the processing imposes on the voices. While, this study does not cover this in detail, it does differ compared to earlier work [2] as the length of the analysis windows is constrained such that delay through filter bank remains below 20 ms. Also the time to compute FFTs and find the optimal representation of the incoming signal adds to the total delay. However, while the faster processors can reduce the delay of the FFT and optimization, the delay arising from constructing the input remains constant regardless of processor.

3. Experiments

Measuring the potential benefits of the known voices separation algorithm involved 13 people with hearing loss and the recently developed Competing Voices Test [3]. In the Competing Voices Test, the listener hears two competing voices before learning which one to repeat to the test leader. It is a taxing test, where good scores require that the listener separate the voices and stores as much separate information about the two sentences as possible. Individual amplification was prescribed with CAMeq formula from the individual audiograms[4], head related impulse responses from the CIPIC database[5] was used to simulate the different spatial configurations over headphones. The speech material was the Danish HINT corpus[6] extended with a recording of the Danish HINT sentences spoken by a female voice and matched in amplitude to the original recordings of the male voice.

The listening test consisted of three conditions shown in Figure 2: two conditions (1+2) presenting two clean voice signals at ±5 degrees and ±45 degrees, and 3) presenting the two estimated voices at ±45 degrees. Condition 1) is the baseline, and the speech separation algorithm worked on these signals. Condition 2) is the upper limit, representing what the speech separation algorithm could achieve if the separation was perfect. Finally, condition 3) measures the performance of the voice separation algorithm shown in Figure 3.

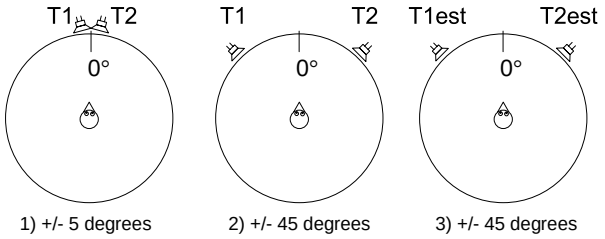

Figure 2: The three test conditions. All source positions and spatial cues simulated over headphones using CIPIC impulse responses. T1/T2 denotes a clean speech signal whereas T1est and T2est denotes the estimated speech signal separated out of a single channel mixture of T1 and T2.

In order to separate the two Danish HINT voices for condition 3, the Known Voices algorithms was given the recordings of the HINT training lists for the male and female talkers and used that to generate the collection of the voice models indicated in Figure 3. This procedure leads to the two estimated speech signals T1est and T2est.

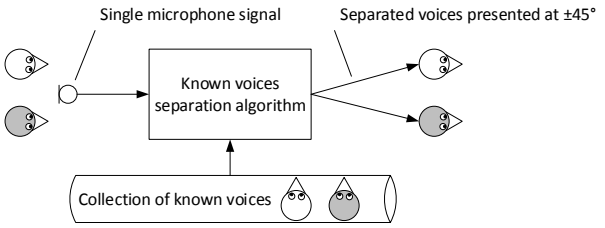

Figure 3: Processing a single mixture of two known voices to produce the estimates of the two voice signals.

4. Results

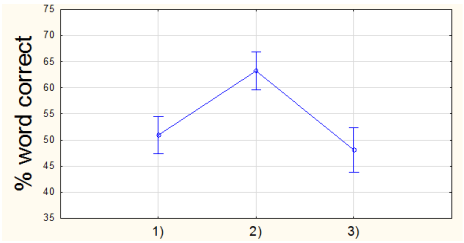

Figure 4: Mean word recognition score in the three conditions. Vertical bars denote 0.95 confidence intervals on the mean.

The results shown in Figure 4 shows the people with hearing loss benefit from the spatial separation of the two voice signals in the competing voices test. The 12% increase is highly significant ($p<0.001$) and in accordance with earlier results obtained with the competing voices test [3]. However, the Known Voices Separation algorithm (condition 3) did not improve speech recognition score for people with moderate sloping hearing loss speech in the competing voices test. However, in verbal feedback, of the test persons reported hearing the outputs of the known voices separation as spatially separated, thus some separation did occur.

5. Discussion and conclusion

While, the present version of known voices separation did not improve the competing voices situation for people with hearing loss, there are several options, e.g., increasing the model size and the maximal latency that have not been explored in depth yet. Relaxing the constraints on model size and maximal latency could enable better separation quality and perhaps better performance. However, at the same time, it would also affect the applicability of the algorithm for many listening situations, as latencies above 20 ms is known to affect lip reading that people with hearing loss also rely on.

In the light of the promising verbal feedback, more research is required to understand why, how, and if the known voices separation will improve. More specifically, the interaction between perceived separation and speech recognition deserves a closer look. The fact that our earlier investigations of Known Voices algorithm demonstrated 2-8 dB attenuation of the competing voice [1] is intriguing in the light of the speech recognition scores and the perceived separation of the two separated voices. There could be two different mechanisms in play here, one that forms the separate objects the streaming, and one that forms the meaning of those separate objects, and that the results show that the threshold for successfully restoring the ability to separate and to recognize speech are different. If this is true, then it also suggests that future testing of speech separation algorithms also require speech recognition tests, as the obtained separation measured in dB is not sufficient to predict speech recognition scores. Furthermore, such extended tests could also enable development of new prediction models that predict speech recognition scores from objective measures of speech separation.

6. References

[1] T. Barker, T. Virtanen, and N. H. Pontoppidan, "Low-Latency Sound-Source-Separation Using Non-Negative Matrix Factorisation with Coupled Analysis and Synthesis Dictionaries," in *40th IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'15*, 2015.

[2] S. Roweis, "One microphone source separation," in NIPS, 2000, pp. 793–799.

[3] L. Bramsløw, M. Vatti, R. K. Hietkamp, and N. H. Pontoppidan, "Binaural speech recognition for normal-hearing and hearing- impaired listeners in a competing voice test," in *Speech in Noise 2015*, 2015.

[4] B. C. J. Moore and B. R. Glasberg, "Use of a loudness model for hearing-aid fitting. I. Linear hearing aids.," *Br. J. Audiol.*, vol. 32, no. 5, pp. 317–325, Oct. 1998.

[5] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[6] J. B. Nielsen and T. Dau, "The Danish hearing in noise test.," *Int. J. Audiol.*, vol. 50, no. 3, pp. 202–8, Mar. 2011.