Algorithms on Noisy Speech for Hearing-Aid Users

Nicholas Prior, Srikanth Vishnubhotla, Carol Espy-Wilson

***Abstract*— The Speech Communication Lab has developed a speech extraction algorithm that is able to remove noise from speech signals, even when the noise is the speech of a competing speaker. The algorithm was previously tested with normal hearing listeners and yielded positive results. Now the algorithm will be tested for hearing impaired listeners to see if it increases the intelligibility of noisy speech for hearing-aid users. Because of the properties of hearing loss and hearing-aids, the tests may not produce the same results for hearing impaired listeners as they did for normal hearing listeners. The algorithm will also be analyzed to see how well it preserves the properties of the original speech signal. The results of these tests and analyses will be put together to see if this algorithm has a potential use in hearing-aids.**

***Index Terms*—Speech extraction, speech segregation, hearing-aids**

I. INTRODUCTION

Hearing-aids have undergone many improvements and modifications since their creation, but their effectiveness in helping the hearing-impaired understand speech in noisy environments still has much room for improvement Even a little bit of background noise gets amplified along with the rest of the speech, making it more difficult to understand the desired speaker.[1] Most current hearing-aids now utilize a dual microphone system, which uses an additional directional microphone to amplify sounds in the direction which the person is facing, while attenuating sounds from other directions.[2] This can be helpful in noisy environments such as restaurants, where the listener is usually facing the intended speaker with the background noise coming from the side and behind. However this method is not always useful in other situations, and there is still much room for improvement in helping hearing-aid users better understand speech in noisy environments.

The Speech Communication Lab has developed an algorithm that is able to reduce background noise in speech signals, even when that noise is the speech of a competing speaker. The algorithm utilizes the properties of speech to separate out unwanted noise, while preserving and amplifying certain regions of the target speaker to improve the overall intelligibility.

The algorithm was previously tested on normal hearing listeners with positive results (Figure 1). The algorithm produced a significant increase in intelligibility of the noisy signals (from red to green) in low signal-to-noise levels, and just surpassed the noisy signals in the highest signal-to-noise level examined. (A full explanation of the testing process and methods is described in the "Listening Tests" section).

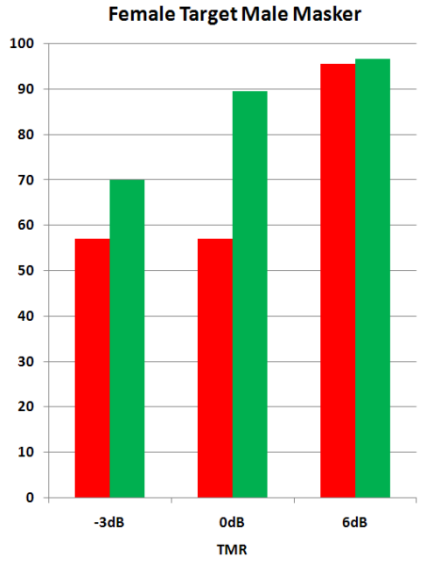

Figure 1. Results of listening tests with normal hearing listeners. Bars represent the percent of keywords correctly identified by the listener; red: noisy signals, green: signals processed by algorithm.

In this study, the effectiveness of this algorithm in improving the intelligibility of noisy speech for hearing-aid users will be tested and analyzed. These tests will see if similar positive results can be achieved for hearing-aid users, and whether or not this algorithm has potential to be used in hearing-aids to improve the intelligibility of speech.

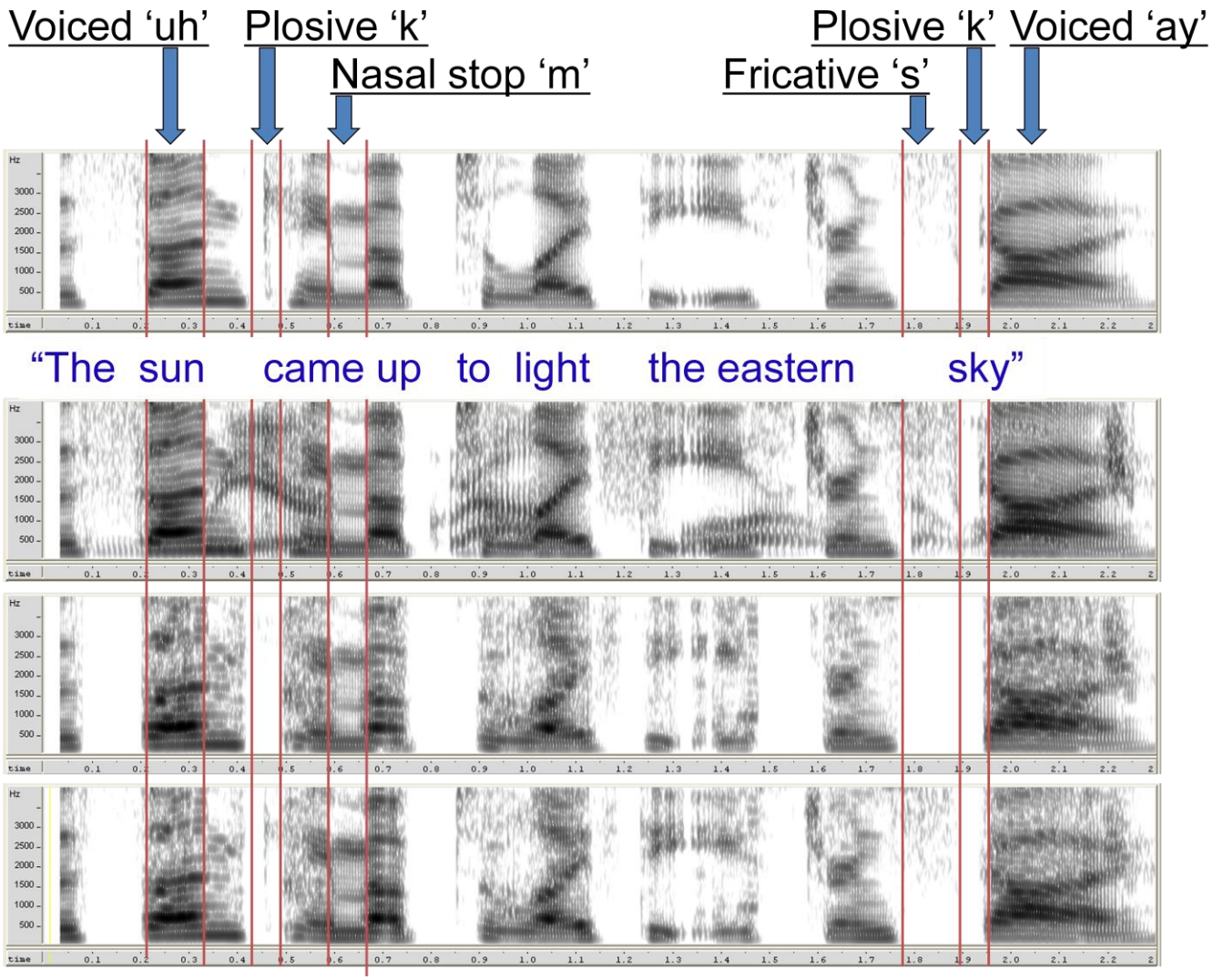

Figure 2. Spectrograms of four different types of the speech signal "The sun came up to light the eastern sky." (1) Original clean speech signal. (2) Noisy corrupted signal. (3) Processed signal. (4) Processed ideal signal.

II. SPEECH PROPERTIES

A. Types of Speech

There are two main types of speech sounds: voiced speech and unvoiced speech. Voiced speech is generated by rapid opening and closing of the vocal cords, which produces a periodic waveform. Voiced sounds include vowels, semivowels, nasals, and voiced consonants. The formants of voiced speech at lower frequencies are generally stronger than the formants at higher frequencies. Unvoiced speech is generated through various formations of the lips, tongue, and teeth, which can constrict or block the airflow to produce speech sounds. Unlike voiced speech, unvoiced speech has an aperiodic waveform, and the formants at higher frequencies are stronger than the formants at lower frequencies. Most consonants are unvoiced. [3]

Figure 2 shows the spectrograms for four different versions of the signal “The sun came up to light the eastern sky.” The first is simply the original clean speech signal; the second is a noisy version in which the clean signal has been corrupted with a background speaker; the third is a processed version generated from the algorithm; and the fourth is an ideally processed version in which correct consonant information has been provided.

Various types of speech have been labeled in the spectrograms. The voiced ‘uh’ and ‘ay’ vowel regions from the words ‘sun’ and ‘sky’ are periodic regions whose formants can be seen as the dark horizontal bands. The plosive ‘k’ region is produced from a blockage of airflow by the tongue, followed by a quick release. The nasal stop ‘n’ region is produced by a blockage of airflow in the mouth by the tongue, while air is able to escape through the nasal cavity. The fricative

's' region is characterized by a constriction of airflow by the tongue, producing a hissing sound.[3]

For these consonant sounds, there are two important physical properties that affect perception: the transient or frication noise produced by the constrictions of the mouth, teeth, and/or lips, and the formant transitions of the neighboring voiced region. Both of these properties make up the consonant sounds of speech.

B. The Algorithm

The algorithm extracts the voiced and unvoiced regions of the speech signal, and selectively amplifies or reduces these regions to try and isolate the target speaker. The two regions that make up the consonant sounds in particular are amplified to better accentuate the consonant sounds of the speaker.

III. LISTENING TESTS

To test the effectiveness of the algorithm, hearing impaired subjects were presented with a series of sentences to listen to and repeat as best they could. Their responses were scored for accuracy based on the number of keywords they were able to correctly repeat for each sentence.

There were four main types of signals presented to listeners. The first type was original clean signals consisting of a female target speaker. The second type was noisy corrupted signals consisting of the clean signals corrupted with a male masker speaker talking at the same time. The third type was processed signals consisting of the noisy signals run through the algorithm to try and remove the masking speaker and preserve just the target speaker. The fourth and final type was ideally processed signals, which were similar to the third type except that the correct consonant information for the target speaker was made available. This fourth type produced a ceiling for the algorithm's potential performance.

The volume ratio of the target and masker speakers (target-to-masker ratio, TMR) was varied for each signal type to simulate different levels of background noise. The speakers were combined at 0 dB (both target and masker talking at same volume level), 6 dB, and 12 dB (target talking much louder than masker).

For the tests, the listeners were presented with a mix of all four signal types, with each type containing a mix of all three target-to-masker ratios. The clean signals provided a control case to verify that listeners could understand the original clean speech.

IV. RESULTS

The results of the tests were not entirely consistent,

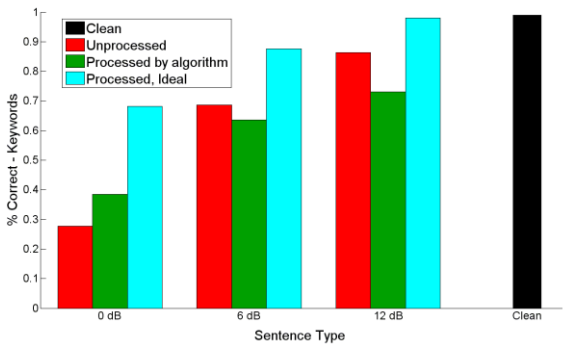

Figure 3. Plot showing the combined results of the listening tests. The bars are grouped by target-to-masker ratio (0dB, 6dB, 12dB).

but they were encouraging. At 0dB TMR, the processed sentences outperformed the unprocessed noisy sentences. However at 6dB and 12dB TMRs, the processed sentences started to fall below the noisy unprocessed sentences, as the target speaker grew naturally louder than the masker speaker. The listeners understood the original clean sentences at a near perfect rate, so all of the listeners could successfully understand normal clean speech.

For all TMRs, the ideally processed sentences were significantly higher than both the noisy unprocessed and the regular processed sentences. This presents the best case performance of the algorithm, and shows that significant improvement in intelligibility can potentially be achieved. If the consonant information of just the target speaker can be better isolated, the algorithm could reach this high level of performance.

While the ideally processed sentences do produce improvement in intelligibility from the addition of the true consonant information, the results also show that the voiced regions of the signal play an important role as well. The algorithm extracts both the voiced and unvoiced regions of the speech signal, but when the true consonant information of the target speaker is given, the intelligibility jumps from green to blue. The remaining difference from blue to black could be partially attributed to missing voiced information of the target speaker that was not recovered by the algorithm.

Normal hearing listeners (Figure 1) showed a greater increase in intelligibility at both 0dB and 6dB than hearing impaired listeners. This could imply that hearing impaired listeners need more emphasis of the consonant information as well as additional voiced information to understand certain words.

While the current performance of the algorithm does not consistently increase intelligibility of noisy sentences, it does have the potential to achieve significant improvement. Further analysis and

experimentation is needed to try and better isolate and amplify the consonant regions of the target speaker, and to achieve a more natural sounding processed signal.

V. ANALYSIS

Initial analysis of the signals has begun to explore and analyze the preservation of important physical properties in the signals. In particular, the energy distribution of the signals is displayed to look at regions of high energy content within the signals. The signals are passed through a filter bank designed to simulate the response of the human ear. The filter bank measures the frequency response of the signal in various frequency bands, and the energy content in each region is calculated. The resulting energy distribution is plotted to generate a 3-dimensional representation of the energy content in the signal with respect to time (x-axis) and frequency (y-axis).

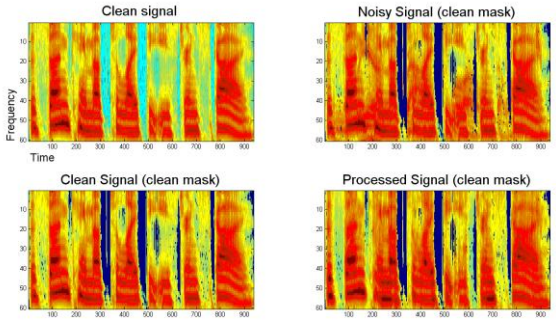

Figure 4. Plots of energy content in the signals with respect to time (x-axis) and frequency (y-axis). Red regions correspond to higher energy content, while blue regions correspond to lower energy content.

Figure 4 shows the energy distributions of the original clean signal, as well as masked versions of the corresponding clean signal, noisy signal, and processed signal. The mask is calculated by specifying an energy threshold for the amount of energy to preserve in the plot. In this case, the upper 90% of the energy in the clean signal is preserved, while the lower 10% of the energy is zeroed out (dark blue regions). This resulting mask from the clean signal is then applied to the noisy and processed signals.

From this mask, we can observe how the noisy and processed signals behave in regions where the clean signal has high energy content. We can also further explore how the formants have been preserved in these regions.

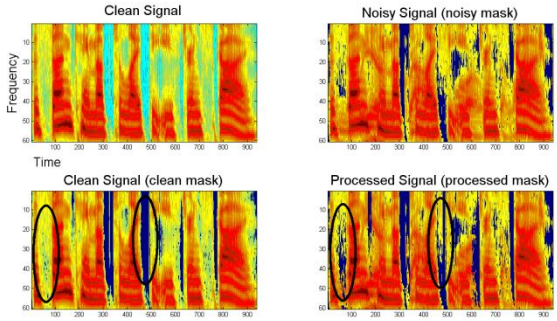

Figure 5. Plots of energy content in the signals with respect to time and frequency. Similar to figure 4, except that the masks are calculated and applied individually for each signal.

Figure 5 is similar to Figure 4 in that it also shows the energy distributions of the signals, except that instead of applying the same mask of the clean signal to the noisy and processed signals, the masks for the noisy and processed signals are calculated separately and are applied individually. By applying each signal with its own mask, we are able to see if any noise or artifacts have leaked into low energy regions of the processed signal, creating more energy where there should be very little. We can also see where the processed signal does not preserve or accentuate certain important information from the original clean signal.

Two regions of interest have been encircled in the clean and processed signals of Figure 5. The first region shows an area where the processed signal failed to reproduce information from the clean signal. The second region shows an area where noise leaked into the processed signal where there should have been low energy content. Once these types of regions have been isolated and identified, further analysis can begin to explore why this is happening and how it can be improved.

VI. CONCLUSION & FUTURE WORK

In conclusion, while the algorithm did not show a consistent increase in intelligibility across all decibel levels, it did show an increase at 0dB TMR. Furthermore, the ideally processed signals showed a significant increase in intelligibility across all TMRs, which demonstrates the potential for improvement and success for the algorithm.

Future work will involve further analyzing and comparing the processed signals with the original clean signals to examine how well the algorithm preserves certain properties of the signal. The energy distributions of the signals will continue to be viewed to try and find patterns and isolate the shortcomings of the algorithm.

In addition to looking at the energy distributions of the signals, PESQ scores will be calculated and compared.

PESQ is a quantitative method for calculating the similarity between two signals.[4] These scores will be calculated for the clean signals and their processed counterparts to try and quantitatively measure their overall similarity.

This future analysis will hopefully result in improvements and modifications to the algorithm to increase the intelligibility of noisy speech.

VII. ACKNOWLEDGEMENTS

This research was funded by the National Science Foundation CISE award #0755224. Additional thanks to the MERIT-BIEN Program, Carol Espy-Wilson and Srikanth Vishnubhotla for this opportunity and the knowledge gained during the program.

REFERENCES

[1] Schaub, Arthur. *Digital Hearing Aids*. New York: Thieme, 2008. Print.

[2] Gnewikow, David, Todd Ricketts, Gene W. Bratt, and Laura C. Mutchler. "Real-world Benefit from Directional Microphone Hearing Aids." *Journal of Rehabilitation Research and Development* 46.5 (2009): 603-18. Web. <http://www.rehab.research.va.gov/jour/09/46/5/gnewikow.html>.

[3] Denes, Peter B., and Elliot N. Pinson. *The Speech Chain: the Physics and Biology of Spoken Language*. New York, NY: W.H. Freeman, 1993. Print.

[4] *Pesq.org - the New Web Portal for Advanced Voice Quality Testing in Telecommunications*. Web. 01 Aug. 2010. <http://www.pesq.org/>.