Passive Amplification to Improve Automatic Speech Recognition

Dean R. G. Anderson, Daniel J. Anderson, Dean G. Anderson, M.D.

ABSTRACT- Passive amplifiers provide additional speech information to improve automatic speech recognition (ASR) in artificial intelligence (AI) applications.

Automatic speech recognition (ASR) begins at the microphone. Is there a way to improve the speech signal before it reaches the microphone? In this paper we apply a new technology, passive amplification, to improve and condition the speech signal prior to microphone detection.[1]

A microphone's performance is impacted by a number of factors including: (1) the distance to the speaker's lips; (2) the speaker's vocal effort; (3) environmental noise; and, (4) microphone equivalent input noise (EIN). The effect of these limitations is a function of frequency. The use of passive amplification, as described herein, increases the signal-to-noise ratio (SNR) for higher speech frequencies and therefore reduces the negative effects of these limitations for higher frequencies.

In previous studies we established the efficacy of Pixation's passive amplification systems using a MEMS microphone for improving speech intelligibility at far-field distances.[2] In this study we provide further evidence that additional speech signal information is available from a MEMS microphone using a passive amplifier for both quiet and noisy environments.

Methods

Two audio files were used in testing. The first was a "speech" audio file. The second was a "speech + babble" audio file which was combined and provided through a single speaker.

The "speech" file was a sequential compilation of speech samples of eight individuals (men and women) reading Aesop's Fables (1,284 words).[3] The readings were previously edited using Audacity® noise reduction and loudness normalization (-23 LUFS per EBU R128-2020).[4,5]

The "speech + babble" file mixed babble at 5 dB below the speech file level. The "speech + babble" file was then loudness normalized to -23 LUFS. Babble was an overlay of 10 one-act plays and Brownian noise.[6] Babble was also loudness normalized to -23 LUFS. The Babble power spectrum exhibited a reduction of 7.67 dB/octave from 500 Hz to 4 kHz ("black noise" = 1/(*freq*^2.94)).

Comparison recordings were made in a large, furnished, reverberant, room. Both a speaker (JBL Control 1 Pro) and microphones (InvenSense® ICS-43432) were pole mounted on 16 mm diameter poles, 1.1 meters above a carpeted floor in asymmetric room positions. The speaker and microphones were separated by 8 meters.

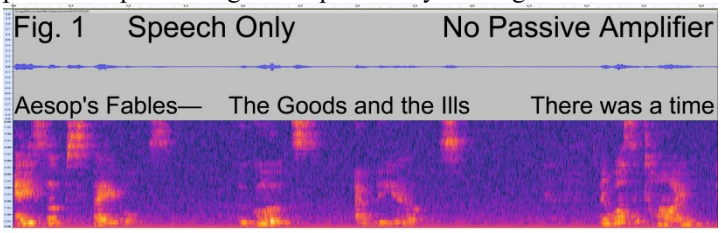
The speaker volume was adjusted to 64.7 dB SPL at 1 meter distance when playing a normalized 1 kHz continuous tone in the reverberant room. Background noise in the room was approximately 37 dB SPL during recordings. Audio levels were determined during a 41.94304 second period using an average of 1,024 sequential measurements.

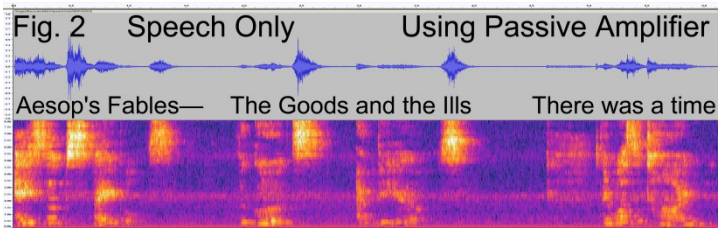
InvenSense ICS-43432 MEMS microphones with I^2S digital audio outputs were used for the comparison recordings. USB digital audio recordings from the I^2S digital audio data were made through an Arduino® compatible Teensy® 4.0 hardware interface with audio sampled at 44.1 kHz.[7] Embedded Teensy programming was used to: (1) attenuate frequencies beyond the speech bands;[8] (2) capture the I^2S output from a first MEMS microphone positioned at the interior apex of a soft silicon 53 mm passive amplifier into the left USB audio channel; (3) capture the I^2S output from a second MEMS microphone with no passive amplifier into the right USB audio channel; and, (4) diminish passive amplifier resonance in the left USB audio channel with a biquadratic notch filter (Fc = 2,380 Hz, Q = 7).

Finite register length effects on rounding/truncation error noise can easily degrade an audio recordings' SNR by 4 dB when using digital signal processing (DSP) near EIN levels.[9] The Teensy 4.0 incorporates an ARM® Cortex®-M7 64-bit floating point math unit for DSP making the effects of finite register length on rounding/truncation error noise for this application inconsequential. It is noted that Pixation Corp. uses fixed-point 32-bit processing with proprietary DSP algorithms to minimize the effects of rounding/truncation noise in Pixation Corp. audio products.

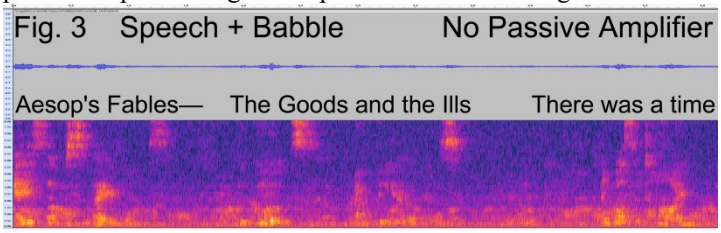
Results

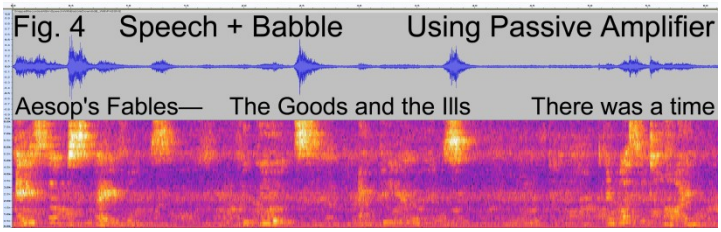
Figures 1 and 2 are used to compare no passive amplifier usage with passive amplifier usage for "speech only" testing at 8 meters.





Figures 3 and 4 are used to compare no passive amplifier usage with passive amplifier usage for "speech + babble" testing at 8 meters.





In Figures 1 through 4, signal waveforms are depicted with the top trace, and spectrograms are depicted in the bottom portion for the

first 6.2 seconds for each recording. All four recordings were post-amplified 34 dB. The spoken speech words are noted.

Figure 5 overlays power spectral densities for no passive amplifier usage and passive amplifier usage for "speech only" recordings.

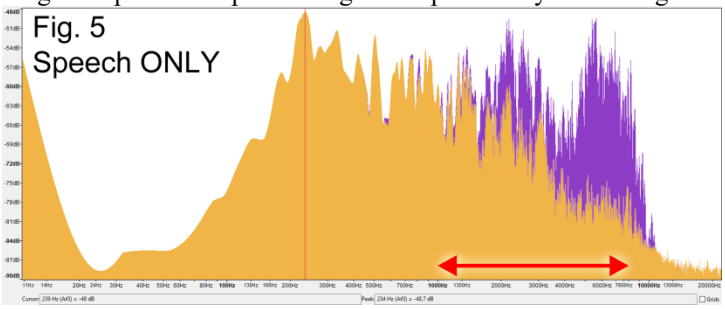
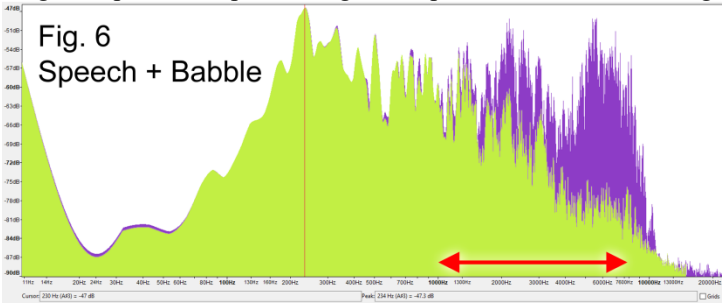


Figure 6 overlays power spectral densities for no passive amplifier usage and passive amplifier usage for "speech + babble" recordings.



In Figures 5 and 6, the purple areas indicate additional speech information extracted by the passive amplifier. The 1 kHz to 8 kHz frequencies are noted between the arrows. The sample periods for power spectral densities were the first 237.8 seconds.

Discussion

The audio SNR in spectrograms is represented visually by color contrast. Phoneme confusion is a function of SNR for both ASR and human speech recognition (HSR).[10] The additional speech information captured by the passive amplifier (shown in Figures 2 and 4) can be used in both ASR and HSR applications to improve phoneme identification.

A power spectral density plot describes the signal's power distribution as a function of frequency when integrated over a long sampling period (e.g., >200 seconds). The overlays of power spectral densities shown in Figures 5 and 6 provide a broad view of where the passive amplifier provides additional signal discrimination benefit. In Figures 5 and 6, passive amplification is shown to provide greater benefit at the higher speech frequencies.

A comparison of Figures 5 and 6 shows that the addition of "babble" noise does not cancel the passive amplification benefit. The "babble" noise used in this study is concentrated in the low frequency bands (below 1 kHz) similar to the noise from working offices, cafés, autos, etc.[11] This distinction is important because: (1) noise is concentrated in low frequencies; and, (2) passive amplification provides increased benefit at high frequencies. This distinction is also important because 70.1% of all speech cues are found in the high ⅓ octave bands (1 kHz to 8 kHz mid-band frequencies).[12] Figures 2, 4, 5 and 6, show that passive amplification increases meaningful information density in higher frequency bands.

As shown in our previous studies, microphone EIN can overwhelm higher frequency bands of speech.[13] A passive amplifier significantly increases meaningful speech information before the microphone EIN is added.

ASR can be improved with large and complex AI models as demonstrated by others.[14] However, large models have high energy consumption, requirements for long sequences of speech, expensive processors, cloud based data transfer with privacy issues, and high word error rates for the initial words spoken, etc.

Convolutional neural networks (CNN) could exploit increased high frequency speech information density from passive amplification. Re-recordings would be required for retraining. These improved CNN applications may enable model size reduction, lower word error rates for brief speech utterances, and lower power consumption for on-device ASR.

Conclusions

Passive amplifiers are shown to extract additional speech information especially in noisy, far-field ASR applications (e.g., 8 meters + speech babble). Additional speech information improves the speech SNR for HSR and may be used to improve ASR accuracy, facilitate immediate response, and enable model size reduction for low-power, on-device ASR.

Contact pixation@pixation.com for additional information. See the pdf version of this document for additional figure detail.

---

[1] *See e.g.*, US Pat. 11,558,690; US Pat. 11,432,066 and other patents at https://www.pixation.com/technology.

[2] Anderson, D. Expanding the Reach of Microphones: Improving Intelligibility. https://www.pixation.com, 2021.

[3] Aesop's Fables (c. 620 BCE - 564) Translated in 1912 by V. S. Vernon Jones (1875 - 1955). https://librivox.org/group/17?primary_key=17&search_category=group&search_page=1&search_form=get_results.

[4] https://www.audacityteam.org/

[5] R 128, European Broadcasting Union. EBU R128-2020 LOUDNESS NORMALISATION AND PERMITTED MAXIMUM LEVEL OF AUDIO SIGNALS. Status: EBU Recommendation, Geneva. https://tech.ebu.ch/docs/r/r128.pdf.

[6] https://librivox.org/one-act-play-collection-012-by-various/

[7] https://www.pjrc.com/store/teensy40.html

[8] ASA Secretariat: Acoustical Society of America. ANSI S3.5-1997 Methods for Calculation of the Speech Intelligibility Index. New York, NY.: Acoustical Society of America; American National Standards Institute, Inc. Approved 6 June 1997: pp.5-6. (62.35 dB SPL at 1 meter is specified as the standard, free-field, speech spectrum level for normal vocal effort.)

[9] Oppenheim, A. Digital Signal Processing. Englewood Cliffs, NJ: Prentice-hall; 1975: pp. 404-479.

[10] Meyer, B. Phoneme Confusions in Human and Automatic Speech Recognition; International Speech and Communication Association; 10.21437/Interspeech. 2007-430https://doi.org/10.1121/1.3458854 https://www.isca-speech.org/archive/pdfs/interspeech_2007/meyer07b_interspeech.pdf

[11] Weisser, A. The ambisonic recordings of typical environments (ARTE) database. Acta Acust. U. Acust. 105, 2019b: pp. 695–713.

[12] ANSI S3.5-1997: pp. 5 (Table 3).

[13] Anderson, D. Understanding Microphone Equivalent Noise. www.pixation.com, 2021.

[14] https://github.com/openai/whisper.