Speech Enhancement for Android (SEA): A Speech Processing Demonstration Tool for Android Based Smart Phones and Tablets

Roger Chappel, Kuldip Paliwal

Signal Processing Laboratory, Griffith University, Queensland, Australia

roger.chappel@griffithuni.edu.au, k.paliwal@griffith.edu.au

Abstract

This paper presents a speech processing platform which can be used to demonstrate and investigate speech enhancement methods. This platform is called *Speech Enhancement for Android* (SEA), and has been developed on the Android operating system, available to students, teaching staff and researchers through the Android market. SEA can be used as an additional teaching tool in undergraduate courses as well as a useful tool to aid researchers in gaining an intuitive understanding of speech enhancement methods. The focus of this platform is to present advanced speech processing concepts in a quick and interactive way on a personal phone or tablet. This paper outlines the operation of SEA along with how it can be used to engage students and speech processing professionals.

Index Terms: Digital signal processing (DSP), analysis–modification–synthesis (AMS), Android operating system.

1. Introduction

Speech Enhancement for Android (SEA) is an educational tool that has been developed to engage students studying speech processing or digital signal processing (DSP) in an attempt to help increase their learning outcomes. It has been developed on the Android operating system with the intention of bringing speech processing principles to the increasingly popular smart phone market. In 2011 Android smart phones accounted for $52.5\ \%$ of the global smart phone market share, double that of 2010 [1]. Additionally, SEA supports a wide range of Android based tablets as they begin to gain momentum in the global market [2].

The objective of this tool is to provide additional flexibility in the delivery of course content to accommodate for shifts in generational learning styles and behaviors [3]. Studies show, that adapting content to suit these trends can encourage and motivate students to engage in activities and enhance their learning outcomes [4].

SEA provides a low cost, interactive and enjoyable tool that brings speech processing principles to the students' personal smart phone or tablet, not only introducing the student to real world applications for DSP such as the field of speech enhancement, but also demonstrating the principles of analysing a human speech signal. These principles address issues such as stationarity, spectral estimation, the role of the magnitude and phase spectrum, statistical methods for noise removal and the spectrogram representation of speech.

The remainder of this paper is organised as follows. The primary elements and functionality of SEA are discussed in Section 2. Section 3 describes how SEA can convey the role of the short-time magnitude and phase spectra on speech intelligibility. Section 4 describes the speech enhancement principles behind the implementation within SEA. Section 5 discusses the educational benefits as a result of a preliminary survey and conclusions are given in Section 6.

2. Primary elements and functionality

SEA allows users to observe four primary elements of speech processing that can aid in the development of an intuitive comprehension of content and allow users to draw connections between theory and real applications. The four elements are:

1. Frame analysis: the short-time analysis of a speech signal both in time and frequency domains.
2. Spectrogram representation of speech: the construction of an image representing phonetic structure of speech across time and frequency.
3. Modification of speech: the role of the magnitude and phase spectrum in terms of speech quality and intelligibility, both with different analysis windows and frame durations.
4. Speech enhancement: enhancing speech which is corrupted by real life noise sources, e.g., train, car, jet, babble, restaurant, street, airport and white noise.

These elements are unified through an interactive user interface (UI), allowing the user to record their own speech, load some pre-packaged examples or load speech from a file as an input. After this they can either add noise of their choosing and commence enhancement, modify speech using one of two different spectral modification techniques or analyse both the short-time magnitude

spectrum and time domain segments of speech. Figure 1 shows a block diagram demonstrating the elements of SEA, from the acquisition of input speech to displaying the spectrogram of both the input speech and enhanced or modified speech.

2.1. Main interfaces

Figure 2 shows three different interfaces the user is presented upon specific touch interactions. Figure 2 (a) is the first interface the user is presented upon opening SEA. There are four primary buttons each assigned to specific tasks. Starting from the bottom left, the "Record" button allows the user to record 3 seconds of speech and upon completion, the recorded speech is transformed into a spectrogram and displayed in the top view. Next is the "Enhance" button; upon a touch event, the enhance button takes user preference settings selected in the configuration menu (see Figure 2 (c)) and performs enhancement accordingly, producing an enhanced audio file and displaying the enhanced spectrogram. To the lower right of Figure 2 (a) is the "Modify" button; upon a touch event, an additional options menu is presented to the user, allowing them to select one of two frequency based modification techniques. Finally, to the top right of the figure is the "Frame analysis" button, that upon a touch event opens a new view (Figure 2 (b)), displaying a frame-by-frame analysis of the spectrogram, allowing the user to scroll through the spectrogram selecting a desired frame to view its time and frequency representations. Figure 2 (c) is the configuration menu, where the user can adjust all configurable settings.
After audio is loaded all associated images can be saved and shared among fellow students or colleagues via Bluetooth, text messages, personal e-mail, or any other supporting service stored on the Android device.

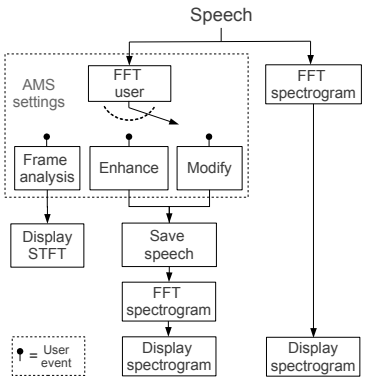

Figure 1: *Block diagram illustrating the functionality of the interactive elements within Speech Enhancement. Each spectrogram is interactive, and upon a touch event can play the associated audio.*

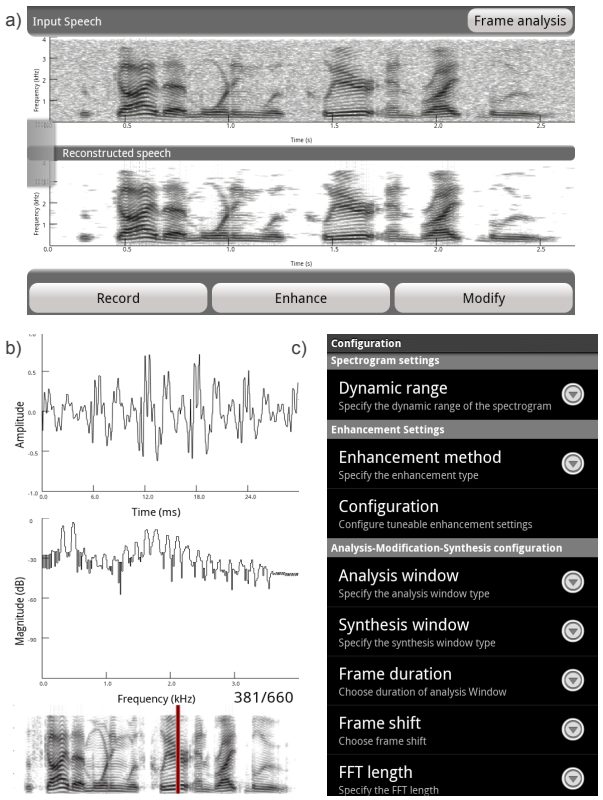

Figure 2: *User Interface (UI): a) The main user interface once speech is loaded and enhanced; b) the frame analysis UI for viewing each frames time and frequency content; here a male speaker is presented with a* 32 *ms window duration,* 4 *ms frame shift and the hamming widow used. The user can seek along the spectrogram to select the desired frame to analyse, and c) the configuration menu, allowing the user to change all configurable settings.*

3. Role of magnitude and phase spectrum in speech

The role of the magnitude and phase spectrum for speech intelligibility has been the topic of many research papers, e.g., [5, 6, 7]. In these studies it has been shown that the magnitude spectrum is known to contribute significantly more to the intelligibility of speech then the phase spectrum when a 20–40 ms Hamming window is employed for speech analysis. However, the phase spectrum has been shown to have a comparable contribution to the intelligibility of speech when a low dynamic range synthesis window, e.g., rectangular window, with $\approx$ 1 s duration is employed for speech analysis. SEA allows the user to observe these findings in a quick intuitive manner by simply changing the analysis–modification–synthesis (AMS) configuration and modifying the speech with one of two modification techniques. The two modification techniques provided are

1. Unit magnitude; and
2. Random phase.

These techniques can be grouped into two categories which are: phase only reconstruction and magnitude only reconstruction. By randomising the phase spectrum, it can be considered as magnitude only reconstruction, whereby the original magnitude is preserved and the original phase spectrum is removed. Similarly, by setting the magnitude spectrum to unity it can be considered as phase only reconstruction, whereby the original phase is preserved and the original magnitude spectrum is removed.

4. Speech enhancement implementation

The primary objective in the field of speech enhancement is to remove noise from a noisy speech signal without causing or introducing distortions in the resulting speech. A number of speech enhancement algorithms have been proposed in the literature to remove noise from noisy speech. These algorithms fall into three categories: 1) spectral-subtractive algorithms, 2) statistical-model-based algorithms and 3) subspace algorithms. In SEA, only AMS based spectral-subtractive and statistical-model-based algorithms were implemented. Spectral subtractive algorithms obtain an estimate of the clean signal by subtracting an estimate of the noise signal from the noisy speech, illustrated in Section 4.1, while, statistical-model-based algorithms aim to estimate a set of discrete Fourier transform (DFT) coefficients that represent the clean speech when only measurements of the noisy speech are observed, illustrated in Section 4.2.

The performance of speech enhancement algorithms vary depending on the signal to noise ratio (SNR) and statistical properties of additive noise. For all statistical-model-based algorithms the noise estimate in SEA is generated by a Voice Activity Detector (VAD) [8].

4.1. Spectral subtraction algorithm

In SEA, one spectral subtraction algorithm has been implemented. Here, we assume an additive noise model: $y(n) = x(n) + d(n)$, where $x(n)$ and $d(n)$ are the clean and noise signals, respectively, $y(n)$ is the noisy signal and n is the discrete time index. By taking the Short Time Fourier Transform (STFT), the additive noise model can be written in complex Fourier notation as

$$Y(i,k) = X(i,k) + \widehat{D}(i,k), \tag{1}$$

where i refers to the frame index and k is the discrete frequency index. In a practical scenario the noise signal is not a known quantity so this model uses the noise estimate $\widehat{D}(i,k)$ which, in the case of spectral subtracting is taken as the average of the first six overlapping frames of speech. Simple transposition of Eq. (1) can be done to extract a clean speech estimate

$$\widehat{X}(i,k) = Y(i,k) - \widehat{D}(i,k). \tag{2}$$

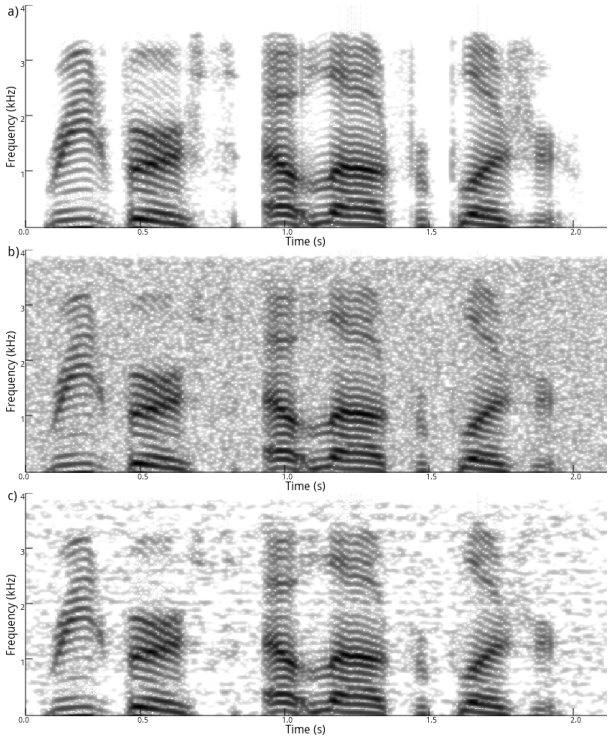

Figure 3: *Spectral subtraction method of speech enhancement; a) clean speech, b) speech corrupted with white noise at* 15 *dB and c) enhanced speech. Image generated by SEA.*

In practice, estimation is performed on the magnitude spectrum only, such that the complex Fourier representation of the enhanced speech becomes

$$\widehat{X}(i,k) = [|Y(i,k)| - |\widehat{D}(i,k)|]e^{j\angle Y(i,k)}, \tag{3}$$

where $\angle Y(i,k)$ is the noisy phase spectrum. Seen in Eq. (3) the resulting enhanced magnitude spectrum ($|\widehat{X}(i,k)| = |Y(i,k)| - |\widehat{D}(i,k)|$) can become a negative quantity if $|\widehat{D}(i,k)| > |Y(i,k)|$, to prevent this a basic half-wave-rectification process was implemented as follows

$$|\widehat{X}(i,k)| = \begin{cases} |Y(i,k)| - |\widehat{D}(i,k)|, & \text{if } |Y(i,k)| > |\widehat{D}(i,k)| \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

A well known shortcoming of this procedure is that inaccuracies in the noise estimate can result in erratic spectral peaks in each frame, which, after signal reconstruction, result in tonal residual noise known as musical noise [9]. The effects of spectral subtraction can be seen in Figure 3, where white noise was added to a clean signal at 15 dB (shown in b) and enhanced using a frame duration of 32 ms with a 4 ms shift and the hamming window as the analysis window. Figure 3 (c) shows the enhanced speech, where the remaining noise is musical in nature.

4.2. Statistical-model-based algorithms

Several other speech enhancement algorithms have been proposed in the literature based off measured statistical properties of the observed noisy signal. These methods make some underlying assumptions in order to derive an estimate of the clean DFT coefficients. Firstly, the additive noise model presented in Eq. (1) is again assumed. Secondly, the STFT expansion coefficients $X(i,k)$ and $D(i,k)$ are assumed to be independent complex zero-mean Gaussian variables, with the expected power $\lambda_x(k) = E[|X(i,k)|^2]$ and $\lambda_d(k) = E[|D(i,k)|^2]$, where $E[\cdot]$ is the expectation operator. A detailed description of these assumptions can be found in [10]. For typical AMS based speech enhancement, an estimate of the clean magnitude ($|\widehat{X}(i,k)|$) is obtained from the noisy signal then reconstructed with the noisy phase as follows:

$$\widehat{X}(i,k) = |\widehat{X}(i,k)|e^{j\angle Y(i,k)} = Y(i,k) \cdot G(k), \qquad (5)$$

where

$$G(k) = \frac{|\widehat{X}(i,k)|}{|Y(i,k)|} \qquad (6)$$

is the spectral amplitude suppression function otherwise known as the gain function for the speech enhancement system. Gain functions are designed to attenuate specific frequencies of the noisy magnitude spectrum to estimate the clean magnitude spectrum. Several gain functions have been implemented in SEA, each with their own attenuation spectral shape. These gain functions are spectral Wiener (SW)[11], Minimum Mean Square Error (MMSE) spectral amplitude (SA) [10], MMSE log-spectral amplitude (LSA) [12] and spectral energy (SE) and are provided below:

$$G_{SW}(k) = \frac{\xi_k}{1+\xi_k}, \qquad (7)$$

$$G_{SA}(k) = \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\exp\left(\frac{-\nu_k}{2}\right)\cdot\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right], \qquad (8)$$

$$G_{LSA}(k) = \frac{\xi_k}{1+\xi_k}\exp\left(\frac{1}{2}\int_{\nu_k}^{\infty}\frac{\exp(-t)}{t}dt\right), \qquad (9)$$

$$G_{SE}(k) = \frac{\xi_k}{1+\xi_k}\sqrt{1+\frac{1}{\nu_k}}, \qquad (10)$$

where

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k. \qquad (11)$$

Where $I_0(\cdot)$ and $I_1(\cdot)$ are given as the zeroth and first order modified bessel functions, respectively. The parameters ξ_k and γ_k are the *a priori* and *a posteriori* SNRs. γ_k is calculated by

$$\gamma_k = \frac{|Y(i,k)|^2}{\lambda_d(k)}, \qquad (12)$$

And ξ_k is calculated using the decision directed method [10].

5. Evaluation

An in-depth pedagogical evaluation on SEA is currently being conducted, however, preliminary findings after surveying ten signal processing research students indicated that 70% of the sample agreed that bringing content to their own personal device increased motivation and 80% agreed that the tool was enjoyable and easy to use.

6. Conclusion

This paper presented a useful tool to demonstrate some foundational and advanced DSP and speech processing techniques for a classroom or laboratory environment. *Speech Enhancement for Android* (SEA) can be placed into the hands of every student or lecturer for the cost of basic stationary. It demonstrates speech processing principles through an interactive and enjoyable process which shows potential in maximizing learning outcomes for students.

7. References

[1] "Market share: Mobile communication devices by region and country, 3q11," Egham, UK, Nov 2011, Gartner Inc.

[2] "Forecast: Media tablets by operating system, worldwide, 2008-2015, 3q11 update," Egham, UK, Sep 2011, Gartner Inc.

[3] E. Rosenbaum and J. A. Rochford, "Generational patterns in academic performance: The variable effects of attitudes and social capital," *Social Science Research*, vol. 37, no. 1, pp. 350 – 372, 2008.

[4] H. O'Neil Jr and R. Perez, *Technology Applications in Education: A Learning View*. USA: Lawrence Erlbaum Associates, Inc., 2003.

[5] H. Hermann von Helmholtz, *On the sensations of tone as a physiological basis for the theory of music*, English ed. Dover, New York: Longmans, Green, London, 1954.

[6] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug 1982.

[7] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. European Conf. Speech Commun. and Technology (EUROSPEECH)*, Geneva, Switzerland, Sep 2003, pp. 2117–2120.

[8] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[9] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 4, Washington, DC, USA, Apr 1979, pp. 208–211.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec 1984.

[11] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: Taylor and Francis, 2007.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr 1985.