ROBUSTNESS IMPROVEMENT OF ULTRASOUND-BASED SENSOR SYSTEMS FOR SPEECH COMMUNICATION

Nemanja Cvijanović[1*], *Patrick Kechichian*[1], *Kees Janse*[1], *Armin Kohlrausch*[1,2]

[1]Philips Research Laboratories, Eindhoven, The Netherlands
[2]Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands
{nemanja.cvijanovic, patrick.kechichian, kees.janse, armin.kohlrausch}@philips.com

ABSTRACT

In recent years, auxiliary sensors have been employed to improve the robustness of emerging hands-free speech communication systems based on air-conduction microphones, especially in low signal-to-noise-ratio environments. One such sensor, based on ultrasound, captures articulatory movement information during speech production and has been used in a voice activity detector and also shown to improve the performance of speech recognizers. However, studies thus far have tested such sensors in ideal scenarios where only relevant articulatory information was assumed to be present. Therefore, in this paper the robustness of such sensors in realistic scenarios is investigated. Challenges arising from non-articulatory movements and other environmental influences captured by ultrasound sensors are discussed and strategies for their detection presented. Finally, the proposed strategies are evaluated in an ultrasound-based voice activity detector.

Index Terms— Ultrasound, articulation, robustness, voice activity detection, Doppler shift

1. INTRODUCTION

Speech communication systems traditionally use one or more air-conduction (AC) microphones to capture and process speech signals, e.g. for speech enhancement. Such systems, however, are susceptible to background noise and reverberation which corrupt the captured speech signal, reducing intelligibility and listener comfort. While a number of methods have been developed to address this problem and enhance the degraded speech signal [1], their performance depends on the application and environment in which they are used, and a general solution to this problem does not exist. This is especially true for low signal-to-noise ratios (SNR) and non-stationary background noise.

Recently, researchers have started employing numerous noise-robust sensors in speech communication systems, such as ultrasonic sensors, which are immune to environmental (audible) noises and reverberation. Unlike AC microphones, when an ultrasonic sensor is aimed at a user's mouth it captures articulatory movement information during speech production which is coded as Doppler shifts in the reflected ultrasound (US) signal. In contrast to sensors such as electromyographs [2] or bone-conduction microphones [3], US sensing is also non-intrusive, i.e. no skin contact is required. Finally, unlike video [4], US sensing does not raise privacy concerns since no sensitive user information is recorded.

The articulatory information provided by US sensors has been successfully employed in speech-based systems. In [5], a US-based lip motion detector was developed to augment the performance of a digit recognition system. A similar sensor was also used in combination with a standard AC microphone for speech recognition systems in [6–9] leading to reductions in word error rate of up to 29 % for a digit recognition task. A speech or voice activity detector (VAD) based on US Doppler sensing was implemented in [10, 11] and it was shown to outperform an AC-based VAD in SNRs below 10 dB. A US-based VAD method which relies on resonance patterns in the US reflection was developed in [12].

In the aforementioned studies, US sensor systems were evaluated under the assumption that only articulatory movements are captured, allowing these systems to leverage the noise-immunity properties of such sensors. However, in practice, this assumption is unrealistic due to user movements and/or other interferences. Thus, the work described here focuses on improving the *robustness* of US Doppler sensing in real-world applications. In particular, we investigate possible events that can corrupt the captured US signal and how these can degrade the captured articulatory information. We then present US features to detect the corresponding artifacts, and implement a US-based VAD as an evaluation framework.

2. ULTRASOUND DOPPLER SENSING

2.1. Articulatory information acquisition

A captured US reflection may contain one or more spectral components outside of the emitted carrier signal's frequency due to the Doppler effect, which describes the change in frequency of a sound wave after being reflected from a moving

[*] This project has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

object. The frequency of this reflection, f_r, is given by

$$f_\mathrm{r} = \frac{c+v}{c-v} f_\mathrm{c} \approx \left(1 + \frac{2v}{c}\right) f_\mathrm{c} = f_\mathrm{c} + \Delta f, \tag{1}$$

where f_c, v, c and Δf denote the frequency of the emitted carrier signal, the velocity of the moving object, the speed of sound, and the Doppler shift resulting from the movement, respectively [13]. The velocity v is assumed to be positive (negative) for movements towards (away from) the sensor. Only movements with vector components parallel to the US beam cause Doppler shifts.

For speech communication systems, the US beam is aimed at the speaker's face during speech production. Under ideal circumstances, Doppler shifts in the reflected signal only contain information about the velocities of various speech articulators such as the lips and cheeks. For this case the reflected signal can be modeled as

$$u_\mathrm{r,id}(t) = \sum_{i=1}^{N_\mathrm{a}} g_\mathrm{a}^i(t) \cos\left\{2\pi\left(f_\mathrm{c} + \Delta f_\mathrm{a}^i(t)\right) + \phi_\mathrm{a}^i(t)\right\} + r_\mathrm{c}(t), \tag{2}$$

where N_a denotes the number of moving articulators and $\Delta f_\mathrm{a}^i(t)$ denotes the Doppler shift resulting from the i-th moving articulator. Associated with the i-th articulator is the phase term $\phi_\mathrm{a}^i(t)$ which depends on the articulator's distance to the sensor and the gain factor $g_\mathrm{a}^i(t)$ which, in addition to the distance from the sensor, depends on its surface area. The term $r_\mathrm{c}(t)$ denotes the reflections of the carrier signal off of fixed objects in the background.

2.2. Extended ultrasound reflection model

In practice, the emitted US beam has a given beamwidth, and is not solely focused on the speaker's face. Therefore, depending on this beamwidth and the speaker's distance to the sensor, any Doppler shifts associated with body movements or movements in the background within the US beam may be captured by the receiver. This is especially true for hands-free applications where the user is sitting at a larger distance from the device. Wideband sound sources which exhibit energy in the US range may also corrupt the received signal. To account for all these non-articulatory contributions, the reflection model in (2) is extended to yield

$$\begin{aligned} u_\mathrm{r}(t) &= \sum_{i=1}^{N_\mathrm{a}} g_\mathrm{a}^i(t) \cos\Big\{2\pi\Big(f_\mathrm{c} + \Delta f_\mathrm{a}^i(t) + \Delta f_\mathrm{bc}(t)\Big) + \phi_\mathrm{a}^i(t)\Big\} \\ &+ \sum_{j=1}^{N_\mathrm{b}} g_\mathrm{b}^j(t) \cos\Big\{2\pi\Big(f_\mathrm{c} + \Delta f_\mathrm{b}^j(t)\Big) + \phi_\mathrm{b}^j(t)\Big\} + r_\mathrm{c}(t) + r_\mathrm{u}(t). \end{aligned} \tag{3}$$

In this model, two types of body movements are distinguished: *connected* and *independent*. *Connected* body movements involve body parts connected to the articulators, e.g., head and torso. When such movements occur during speech activity, the velocity of the body part and that of the i-th articulator are added vectorially to produce a shifted articulator velocity relative to the sensor. These type of movements are represented by the Doppler shift term $\Delta f_\mathrm{bc}(t)$ which further shifts the original Doppler term, $\Delta f_\mathrm{a}^i(t)$, associated with the i-th articulator in (3). *Independent* disturbances represent the movement of objects or body parts which are not directly connected to the articulators. These produce Doppler shifts that are superimposed on the articulators' Doppler components and can irrecoverably mask articulatory information during speech activity.

The second summation term in (3) includes the effects of disturbances associated with independent as well as connected movements. N_b denotes the total number of moving surfaces and $\Delta f_\mathrm{b}^j(t)$ denotes the Doppler shift in the US signal corresponding to movement of the j-th surface. The gain and phase terms $g_\mathrm{b}^j(t)$ and $\phi_\mathrm{b}^j(t)$ are associated with the j-th surface and depend on its area and proximity to the sensor while the last term, $r_\mathrm{u}(t)$, models the leakage of high-frequency sounds in the system environment into the US receiver. Most naturally occurring sounds exhibit a fall off in the high-frequency portion of their spectra due to absorption, especially within the narrow US bandwidth considered here (39 kHz - 41 kHz). However, impulse-like sounds such as knocks or clicks may contain sufficient energy to corrupt the US reflection.

3. FEATURES FOR ROBUST DOPPLER-BASED SPEECH PROCESSING

As discussed in the previous section, different movement types and broadband sounds can corrupt the articulatory movement information contained in a reflected US signal. Current systems that make energy-based VAD decisions or extract Doppler-shift features related to articulation typically rely on analyzing frequency bands around the US carrier frequency. For these systems the presence of such US *noise* can render them unreliable. Therefore, to properly track the desired articulatory information in both time and frequency, a number of features are proposed in this section.

Connected body movements offset the Doppler shifts associated with articulator movements, and additionally shift a significant portion of the emitted US energy in frequency, depending on the connected body parts involved in the movement. To detect and track these frequency shifts, a maximum tracker is defined,

$$k_\mathrm{max}^l = \arg\max_k \left|U^l(k)\right|, \tag{4}$$

where $U^l(k)$ is the FFT of the l-th frame, and k denotes the frequency bin index. For the ideal situation in (2) or in situations where the shifted US carrier energy is still less than the energy of $r_{\mathrm{c}}(t)$ in (3), the output of the tracker is k_{c} corresponding to the original carrier frequency f_c. Unlike other measures such as center of gravity, the maximum tracker provides a more accurate estimate of the carrier shift.

Movement is detected for a given frame if k^l_{max} deviates significantly from the original carrier frequency bin k_{c}. In addition to a Doppler shift, a spread in carrier bandwidth around k^l_{max} can also be observed as a result of certain body movements. This spectral spread is attributed to the fact that the human body is not rigid or perfectly parallel to the US sensor during natural movement, resulting in a non-uniform spatial distribution of velocity components. Therefore, in order to correctly isolate and recover parts of the shifted articulatory information, not only the Doppler shift corresponding to k^l_{max}, but also this spread must be correctly estimated.

Here, the carrier spread is estimated using knowledge about the velocity of the body movement estimated from (4) – the faster the movement, the higher the maximum change, $|k^l_{\text{max}} - k_{\mathrm{c}}|$, and the higher the spread. First, only frequency bins whose energy is sufficiently lower than the maximum energy are included in subsequent analysis steps. This bin threshold is determined by a pre-defined base difference E_{b} and the difference in the total energy in the analyzed frame l and a smoothed version of it, $T^l_{\mathrm{c}} = E_{\mathrm{b}} + E^l_{\text{tot}} - \bar{E}_{\text{tot}}$ (in dB). The comparison to a smoothed version of the total frame energy $\bar{E}_{\text{tot}}$ is utilized to react to changes in the position of the speaker relative to the sensor, e.g., if the speaker moves closer, the total frame energy increases as more of the US signal is reflected off the speaker, increasing T_{c}. Only bins with energy T_{c} dB below the maximum are used for the VAD decision resulting in a bin threshold k^l_{th}. Furthermore, k^l_{th} can be adapted based on the maximum shift as $\hat{k}^l_{\text{th}} = k^l_{\text{th}} + g(|k^l_{\text{max}} - k_{\text{carrier}}|)$, where g is a function that depends on the application and required VAD sensitivity.

The aforementioned scheme might still generate false alarms due to independent body and background movements which produce sufficient energy in a large bandwidth around the carrier frequency – in this case the maximum will not necessarily deviate from k_{c}. Hence, an additional detection strategy based on the spectral symmetry of the US reflection signal is applied. For this measure, the energy in frequencies above and below the carrier are compared for each frame l,

$$s^l = \frac{\min(\sum_{k<k_{\mathrm{c}}-\epsilon} |U^l(k)|, \sum_{k>k_{\mathrm{c}}+\epsilon} |U^l(k)|)}{\max(\sum_{k<k_{\mathrm{c}}-\epsilon} |U^l(k)|, \sum_{k>k_{\mathrm{c}}+\epsilon} |U^l(k)|)}, \quad (5)$$

where $0 \leq s^l \leq 1$ and ϵ is used to determine how many bins around the carrier to ignore in the symmetry calculation. The reason behind this is that the emitted US signal is not perfectly centered at the carrier frequency but smeared onto neighbouring bins, even without any movement. The motivation for the symmetry measure is that for certain FFT window lengths (above 10 ms), regions in the US reflection that correspond to speech exhibit a high level of symmetry around the carrier frequency ($s^l > S_{\text{low}}$), unlike regions corresponding to body and background movements. For impulse-like noise events, the energy is evenly smeared out across the US spectrum and s^l is close to unity. However, even though the US spectrum shows high levels of symmetry during speech activity, the corresponding value of s^l is unlikely to exceed $S_{\text{high}} = 0.9$ in practice, enabling the system to differentiate such noise from speech. These thresholds are set depending on the application. A higher S_{high} enables fast reactions and more accurate detection but might cause misclassifications, as the symmetry can drop during normal articulatory movements. A low threshold, however, may detect artifacts too late, for example after a VAD has already triggered a speech detection decision.

4. EXPERIMENTAL RESULTS

4.1. Setup

An ultrasound sensor consisting of a US receiver (Prowave 400SR100) and a US transmitter (Prowave 400ST100) with a 40 kHz carrier frequency and a $72°$ beamwidth, was constructed. This sensor was aimed at a speaker seated at a distance of 50 cm away and directly facing the sensor during all recordings. The sensor's vertical position corresponded to the height of the speaker's mouth. At this interaction distance, which is typical for most hands-free audio communication devices, the sound pressure level of the transmitted ultrasound signal was measured to be approximately 80 dB (SPL). To reduce signal leakage from the US transmitter into the receiver, the inter-sensor spacing was set to 10 cm.

All recordings were made in a moderately reverberant room ($T_{60} = 400$ ms). A sampling frequency of 96 kHz was used during the recordings while the speaker was reading from a list of Harvard sentences. The received US signal was downmixed to 4 kHz, lowpass filtered with a cut-off at 8 kHz, and resampled at 16 kHz for further processing. The reflections were then analyzed using 64 ms Hann-windowed frames with 50% overlap.

4.2. Evaluation framework

A VAD, or speech detector, is an integral part of many speech processing applications that provides information on whether speech is present or not. We used a US-based VAD similar to that in [10] and [11] to evaluate the proposed features for robustness improvement.

To detect whether speech is present in a given frame, the total energy, $E_{\mathrm{a}}(t)$, in the frequency bins deemed to contain possible articulatory information is compared to a threshold T_{s}. In [10] and [11] an adaptive threshold was defined and used for this purpose. Here, we define a different adaptive threshold, which is based on the the minimum statistics estimator approach in [14]. The minimum articulatory energy over the last L s of the US signal is used as the baseline $E_{\text{min}}(t)$. The threshold is then defined as $T_{\mathrm{s}} = E_{\text{min}}(t) + E_{\mathrm{s}}$,

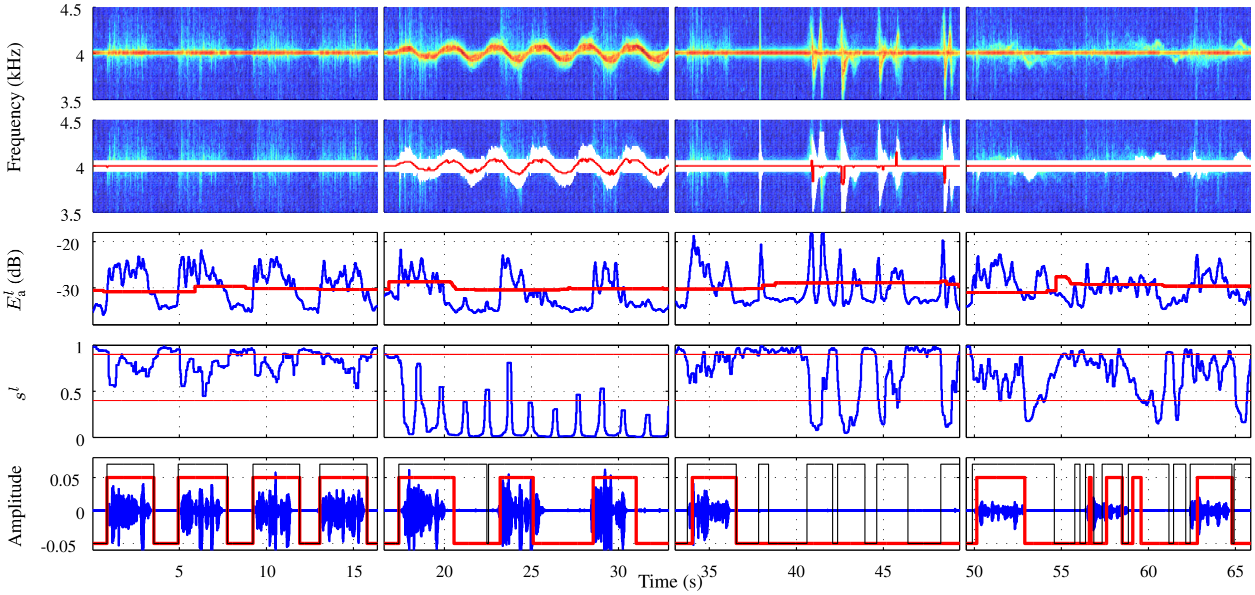

Fig. 1. US-based VAD performance with and without additional robustness features for various disturbance scenarios using data from one speaker. The five rows from top to bottom plot the: US reflection spectrogram; carrier spread detection; energy in the region of interest with the adaptive VAD threshold; symmetry feature value with fixed symmetry thresholds; and the original AC signal, the basic US-based VAD output (thin line) as well as the VAD output with artifact detection (thick line).

where E_s is a fixed, predefined dB value. A speech presence decision is then made in case the articulatory energy in the current frame exceeds T_s. Furthermore, a hangover scheme is applied to prevent mid-speech clipping in case speech is misclassified as silence.

This basic VAD approach, as well as the approaches in [10] and [11] only exhibit a high performance in the ideal case modeled by (2). Therefore, it is expected that the features proposed in Sec. 3 will improve the detection accuracy for various movements and broadband noise. For that purpose, the recording was divided into the four time segments shown in Fig. 1. The first segment corresponds to the ideal scenario where the speaker remains still and the sensor captures only articulatory information. In the second segment, the speaker makes connected upper-body movements by swaying towards and away from the sensor (roughly ± 20 cm). Independent body movements related to the speaker consisting of arm movements in front of the sensor were made in the third segment. Additionally, an impulse sound corresponding to a finger snap was generated at 38 s. Finally, in the last time segment, independent disturbances generated by another person walking in the background of the speaker were recorded.

In the following, the VAD parameters are set to $L = 5$ s and $E_s = 5$ dB. A moving average filter of order 5 is used to smooth the articulatory energy contour and a hangover scheme is applied during the VAD decision in which the output is smoothed over the last 5 frames for speech to silence transitions. The results of the baseline and proposed VAD are shown in the last row of Fig. 1. To quantify the behaviour of the proposed features, Fig. 1 also plots the articulatory energy E_a^l and the symmetry measure s^l in the third and fourth rows, respectively. The articulatory energy in the ideal case and the symmetry measure exclude the first three bins around the carrier with $\epsilon = 3$. The high and low limits for this measure, S_{low} and S_{high} , are set to 0.4 and 0.9, respectively. A frequency range of up to 1 kHz above and below the carrier (3 kHz - 5 kHz for the downmixed signal) is considered to calculate energy values E_{low}^l and E_{high}^l where $E_a^l = \max(E_{\text{low}}^l, E_{\text{high}}^l)$. The second row of Fig. 1 highlights the detected carrier spread in the original US spectrogram which is left out of the articulatory energy computation as well as the output of the maximum tracker defined by (4). For this spread estimation E_b was set to 35 dB.

4.3. Results and discussion

From the VAD decision results in the bottom row of Fig. 1, the proposed robustness features for VAD clearly result in an improved accuracy for the various scenarios in the four time segments. In the ideal situation, with no movements present, both VADs display the same level of performance as expected. For the second column, where the user generates connected body movements, the basic energy based VAD continuously detects speech after the first onset. This is due to the fact that the basic VAD assumes that any energy in frequency bins outside of the original carrier frequency band ($|k - k_c| > \epsilon$) corresponds to articulatory movements. With the proposed maximum tracker and carrier spread estimator, the highlighted regions in the spectrogram displayed in the second row are excluded from the calculation of E_a which leads to a significant reduction in false alarms. Furthermore, a compensation of the movement artifacts and recovery of parts of the articulatory information is possible. However, it is important to note that in some cases a large carrier spread can completely mask articulatory information, which can lead to missed speech detections or clipping in the VAD decision. This can be seen at the end of the second utterance between 22 s and 26 s where the VAD misses the last portion of the sentence. The values of E_s and E_b can be adjusted to increase sensitivity to movements at the cost of an increased number of false alarms.

Robustness to independent movements as well as impulse-like noise is displayed in the third column of Fig. 1. While the energy E_a^l exceeds the adaptive threshold on numerous occasions, the symmetry measure is able to correctly rule out these US artifacts as shown in the fourth row, in particular for the impulse generated around the 38 s mark. This is in contrast to the baseline VAD which detects these movements and noise as speech. In the last column of Fig. 1, independent movements in the background lead to artifacts around the original carrier frequency. Due to the larger distance at which these movements occur, their contributions are difficult to differentiate from articulatory information. This can lead to an increase in the estimated articulatory energy and trigger the VAD to detect speech. Such artifacts can be detected using the symmetry measure which leads to a decreased false alarm rate as shown in the bottom row of Fig. 1.

5. CONCLUSION AND OUTLOOK

In this work, the robustness of a US sensor in a hands-free speech communication system is investigated based on a model of the captured US signal. Practical challenges related to artifacts caused by non-articulatory movements and broadband noise captured by the sensor were discussed and their corresponding artifacts analyzed. A set of features was developed to detect these artifacts, and finally evaluated in a voice activity detection framework. The comparison between US-based VADs with and without their employment in various usage scenarios showed that the implemented features considerably improve robustness and overall accuracy. Although a VAD was used for evaluation purposes, the proposed strategies remain valid and are applicable for many systems incorporating US sensors. Finally, with the proposed methods, future research will look at extracting more complex articulatory features from the US reflection signal and link these to traditional acoustic features extracted from a microphone.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech enhancement*, Signals and Communication Technology. Springer, 2005.

[2] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, pp. 341–353, 2010.

[3] P. Kechichian and S. Srinivasan, "Model-based speech enhancement using a bone-conducted signal," *Journal of the Acoustical Society of America*, vol. 131, pp. 262–267, 2012.

[4] S. E. Hudson and I. Smith, "Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems," in *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, 1996, pp. 248–257.

[5] D. L. Jennings and D. W. Ruck, "Enhancing automatic speech recognition with an ultrasonic lip motion detector," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 1995, pp. 868–871.

[6] S. Srinivasan, B. Raj, and T. Ezzat, "Ultrasonic sensing for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 5102–5105.

[7] Karen Livescu, Bo Zhu, and James Glass, "On the phonetic information in ultrasonic microphone signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4621–4624.

[8] K. Kalgaonkar and B. Raj, "Ultrasonic Doppler sensor for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4865–4868.

[9] B. Zhu, T. J. Hazen, and J. R. Glass, "Multimodal speech recognition with ultrasonic sensors," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH'07)*, 2007, pp. 662–665.

[10] K. Kalgaonkar and B. Raj, "An acoustic Doppler-based front end for hands free spoken user interfaces," in *IEEE Spoken language technology workshop*, 2006, pp. 158–161.

[11] K. Kalgaonkar, R. Hu, and B. Raj, "Ultrasonic Doppler sensor for voice activity detection," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 754–757, 2007.

[12] I. V. McLoughlin, "Super-audible voice activity detection," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 9, pp. 1424–1433, 2014.

[13] M.A. Richards, *Fundamentals of Radar Signal Processing*, chapter 2.6.1, pp. 92–95, McGraw-Hill, 2005.

[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001.