Angular Dependency of Human Speech Recognition using Interferometry Radar

Christopher Williams
Electrical and Computer Engineering
Texas Tech University
Lubbock, USA
christopher.williams@ttu.edu

Changzhi Li
Electrical and Computer Engineering
Texas Tech University
Lubbock, USA
changzhi.li@ttu.edu

2024 IEEE Radio and Wireless Symposium (RWS) | 979-8-3503-4045-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/RWS56914.2024.10438635

***Abstract*— The use of radar in speech recognition is a growing field of interest in recent years. The inclusion of radar introduces several new variables into the scenario. One of the lesser-known variables of this process is the effect that a radar's inherent angular dependency has on the recovered speech signals. This paper presents a study into this effect seen through continuous wave (CW) Doppler radar. The vocal vibrations were detected without aid of secondary passive or active amplification to isolate the radars sensitivity to this effect. The phrases considered in this study were the words "A," "I," and "O." The readings of each phrase were compared across several angles of observation to map the difference in signal-to-noise ratio (SNR) from 80 – 100 degrees for a person.**

Keywords*— *continuous wave (CW) interferometry Doppler radar detection, short time Fourier transform (STFT), signal-to-noise ratio (SNR)

I. Introduction

Since the 1950's speech recognition has been an area of great interest in the scientific community [1]. Speech recognition has been used in various technologies for a variety of reasons. These uses range from interaction with navigation to healthcare systems. In all these instances it is used as a more intuitive way of interacting with technology for the user. Such as using text to speech while driving a car. In recent times a greater interest has been placed on speech detection systems using sensors in conjunction with a conventional microphone. One of the routes explored has been the integration of radars into speech recognition systems [2]. These systems make use of the unique qualities of radars to improve the robustness of the system to environmental factors. The radars in most of these cases were used to both read a person's speech patterns as well as to help complete missing components from the audio reading. By introducing a radar, these systems can improve their resistance to environmental circumstances due to the ability of the waves radiated from a radar to pass through many environmental factors. Radars are also used in these types of systems due to their ability to be very directive in what they detect. This allows for radar to negate more active factors in an environment such as someone speaking loudly in a different direction than the subject of interest.

The addition of radar into these speech recognition systems brings up questions regarding the difference in the nuances of reading speech with a microphone versus a radar. These differences come from the changes in the medium used for detection. For microphones their detection is done through the direct recovery of sound waves propagating through the air. While for radar, they use electromagnetic waves as the method of propagating through mediums. This paper focusses on the effect that the inherent angular dependency of radar has on the effectiveness of reading the vocal vibrations of a target across multiple angles. The area of the body explored in this paper is the human throat during speech. The reduced amount of tissue between the vocal vibration and the surface of the throat makes it ideal for study due to the lowered amount of attenuation to the vibrations generated by the subject's voice box.

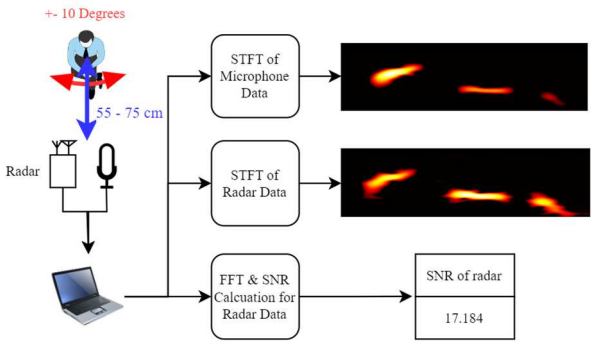

Fig 1: Setup and process for speech detection via radar and microphone.

II. Theory and Signal Processing

A. Background for This Study

To provide the highest sensitivity to detecting vibrations a Doppler radar was used in these experiments. Doppler radars detect movement by transmitting a signal at a constant frequency and reading the reflection from moving objects in the transmission path. The movement of objects in the radar's transmission path induces a change in the received frequency. In the case of this experiment the movement being observed is the radial displacement relative to the radar caused by a person's throat during speech. These changes in frequency, or Doppler shift, allow for the observation of the characteristics of the target's movements. Continuous wave (CW) Doppler radars are commonly used in cases where the motion considered is small due to their higher sensitivity to minor changes in the velocity of an object in comparison with other types of radar. Since the movement under study was the vibration of a subject's throat, micrometers peak to peak, a Doppler radar was

The authors wish to acknowledge National Science Foundation (NSF) for funding support under Grant ECCS-2030094.

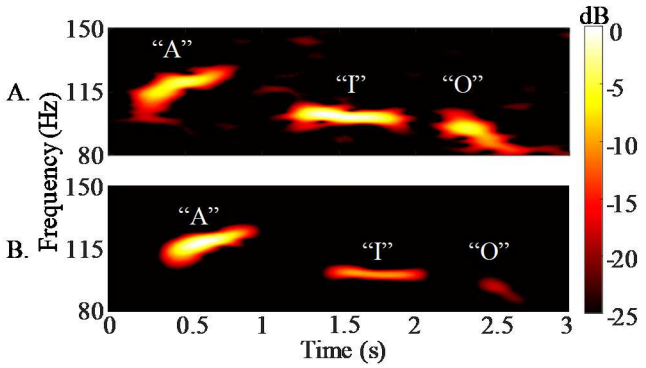

Fig. 2: Comparison of radar readings (A) to microphone readings (B) for 90 degrees relative to the radar at 55 cm away from subject.

used. The response of these motions results in micro-Doppler responses that interferometry Doppler radars are uniquely suited to detect. The scenario considered for this study was that of a subject sitting in front of a radar at various ranges and angles. At set positions and rotations relative to the radar the subjects spoke intermittently for 15 s to create a mapping of the signal-to-noise ratio (SNR) relative to a radar's angle of observation. The phases used for testing were "A," "I," and "O" due to their commonality in the English language. These readings were then compared against microphone readings to verify the capturing of true audio data before further processing. An overview of these tests is shown in Fig. 1. The process behind this will be further discussed in Section II.B. The radar used for these tests is the BGT60LTR11AIP. An off-the-shelf 60 GHz interferometry Doppler radar made by Infineon Technologies. The words were all performed with a 60 Hz metronome running to help make the performance of the words consistent throughout the time frame of the tests. Two subjects were used for these tests. The subjects studied were male, aged 22 – 26 of about the same body type. The age and body types of the subjects were kept consistent to lower the influence differing physiology has on the results. They were both found to not have any respiratory issues prior to the tests.

B. Vocal Vibration Detection and Signal Processing Process

Due to the vibrations made by the throat being equitable to the detection of micrometer and below movements. An interferometry Doppler radar was used to make use of the *I* and *Q* channels to better detect micrometer movements of the targets through differential detection. From [3] the mathematical representation of the *I* and *Q* signals received after the baseband of the radar can be stated as (1):

$$I(t) = \cos\left(4\pi\frac{\Delta d}{\lambda}sin(\omega_{Sp} * t) + \varphi_{Sp}\right),$$
$$Q(t) = \sin\left(4\pi\frac{\Delta d}{\lambda}sin(\omega_{Sp} * t) + \varphi_{Sp}\right). \quad (1)$$

Δd is the magnitude of displacement that the throat experiences during speech. λ is the wavelength of the transmitted signal. φ_{Sp} is the phase shift created through interaction with the subject, and ω_{Sp} is the frequency of the vibrations during speech. ω_{Sp} is the key factor to be extracted due to it containing the speech information of both the fundamental frequency and the harmonics of the words spoken by the subjects.

For the processing of the received radar signals a conventional CW Doppler process was used. First the DC component of the signal read was removed to prevent any DC circuit bias from the radar carrying throughout the signal processing. Then the amplitude of the signal is normalized to generalize the magnitudes of the signal data being considered. The signal is then passed through a high pass filter with a cutoff frequency of 80 Hz. This is to avoid the Doppler frequency response caused by the movement of the subject's jaw during performance as well as any other low frequency noise appearing in the data. The movement of the jaw and other low frequency sources were considered clutter to the desired values. Finally, the *I* and *Q* signals are then combined digitally via complex signal demodulation [4] to create the full signal (2).

$$S(t) = \cos\left(4\pi\frac{\Delta d}{\lambda}sin(\omega_{Sp} * t) + \varphi_{Sp}\right) + i * \sin\left(4\pi\frac{\Delta d}{\lambda}sin(\omega_{Sp} * t) + \varphi_{Sp}\right) \quad (2)$$

The resultant demodulated signal is then run through a Short Time Fourier Transform (STFT) to extract the spectral components of the signal vs time. Using this process, the recovered ω_{SP} can be extracted from the radar signal. Plotting the STFT results allows for the visual confirmation of the contents of the speech patterns and their characteristics before further processing. Using this process, the range of fundamental speech frequencies for the subjects were found to be centered around 120 Hz. Comparing this result against the work done in [5] allowed us to validate the radar's capturing of the fundamental frequencies for the subjects. An example of the recovered STFT's for the radar versus the microphone for subject 1 can be seen in Fig. 2.

C. SNR Calculation

The data processing for the SNR at each angle was done in the frequency domain through the use of the STFT. The SNR of the

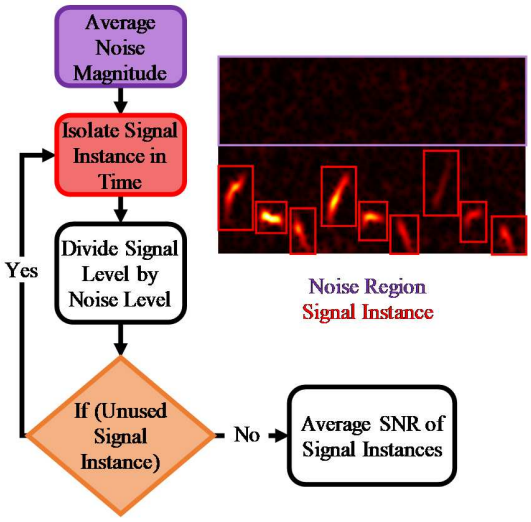

Fig. 3: Block diagram of STFT calculation process with example STFT plot for 90 degrees of subject 1.

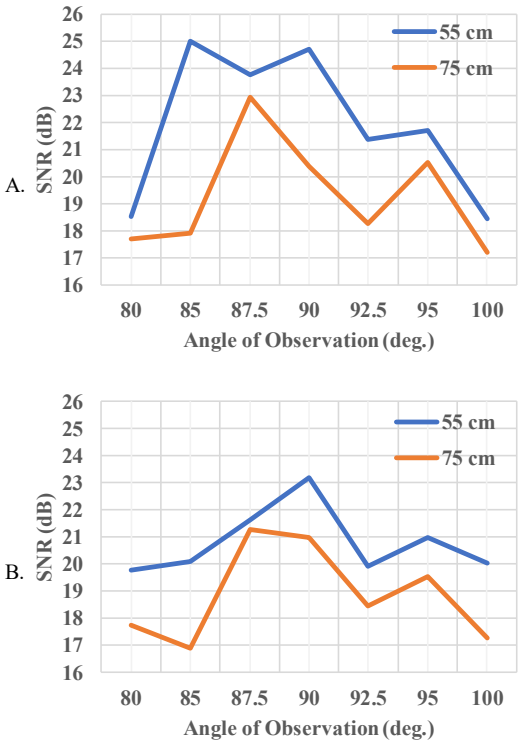

Fig. 4: Graphs of the averaged signal-to-noise ratio (SNR) from radar readings for subject's 1 (A) & 2(B). Readings are for 55 and 75 cm away from subject over 90 +-10 degrees.

signal was derived through isolating the instances of speech in the time domain through the STFT. The noise level of the signal was taken as an average over 9 seconds of the frequency range of 150 – 200 Hz to get a noise floor for the experiment. The noise frequency range was decided through visual confirmation of the signal after the STFT to avoid the harmonics present in the recordings. To get the average signal magnitude for an experimental run, the peak magnitude of each spectral instance from the speech was isolated in time and averaged together. The average of the signal magnitude and noise magnitude were then entered into equation (3) to get the SNR for each angle.

$$SNR_{\angle N} = 20 * Log_{10}\left(\frac{Signal_{avg}}{Noise_{avg}}\right) \quad (3)$$

An overview of the SNR calculation process is shown in Fig. 3.

The angles (N) considered for this study were 80, 85, 87.5, 90, 92.5, 95, and 100 degrees relative to the radar. These readings were then taken at both 55 and 75 cm away from the radar to assess their variance over distance. The outliers in the data caused by experimental error were then removed, and the remaining data at each angle were averaged together in linear scale to create a general reading for the person. This process allowed for a general mapping to be created of the person under test for 90 +- 10 degrees. The results of the two people tested can be seen in Fig. 4.

III. Experimental Setup and Results

It can be seen from Fig. 4 that there is a vast difference between relatively slight changes in angles. An example of this is the difference seen in the jump of ~6.5 dB from 80 – 85 degrees at 55 cm away from the target for subject 1. A consistent trend throughout the data collected for both subject 1 and subject 2 is that the optimal angle to detect a person lies in the 85 – 87.5 degrees range at both 55 and 75 cm. This runs counterintuitive to commonly held assumptions of the optimal angle being 90 degrees relative to the radar. The fall-off in SNR from this point varies from subject to subject with subject 2's SNR being stable throughout 87.5 – 92.5 degrees. The delta between the maximum and the minimum SNR also varied depending on the person. For subject 1 the average difference between their max and min was 6.13 dB. This equates to about a 2x increase in the linear scale. Meanwhile, subject 2 has an average delta in SNR from max to min of 3.9 dB or 1.6x difference in the linear scale. The differences in SNR delta over these ranges can be attributed to the differences in the muscular structure of the subject as well as the directivity of the radar system relative to a microphone.

IV. Conclusion

In this paper the direct non-assisted radar readings for speech detection are studied. Through these readings a mapping for 80 – 100 degrees relative to radar was constructed for 2 subjects. It can be seen from the findings that the optimal detection point for a person lies in the range of 87.5 to 92.5 degrees up to 75 cm away from the subject. Another property discovered through the study is that there was a lot of variance in the derived SNR across each angle of observation relative to the radar. For 55 cm to 75 cm with both subjects the SNR follows a normal distribution with the peak being in the 87.5 – 92.5 degrees region. The drop in SNR for both subjects varied vastly over a slight change in the angle of observation. This shows an inherent issue with using radars for speech detection due to their angular dependency. It can be inferred from this study that the vocal vibration of the human throat is not independent of the angle of observation for the radar. The readings are very dependent on both the muscular structure of the subject's throat and the angle of observation of the sensor in question. This work can be further expanded through the introduction of a larger sample size of subjects as well as the use of finer degree steps during experimentation. In future works this approach can be used in conjunction with machine learning to compensate for the SNR profile of a subject.

References

[1] B. H. Juang and L. R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development".

[2] T. Liu et al., "Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals," in Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, Coimbra Portugal: ACM, Nov. 2021, pp. 97–110.

[3] C.-S. Lin, S.-F. Chang, C.-C. Chang, and C.-C. Lin, "Microwave Human Vocal Vibration Signal Detection Based on Doppler Radar Technology," IEEE Transactions on Microwave Theory and Techniques, vol. 58, no. 8, pp. 2299–2306, Aug. 2010.

[4] C. Li and J. Lin, "Complex signal demodulation and random body movement cancellation techniques for non-contact vital sign detection," 2008 IEEE MTT-S International Microwave Symposium Digest, Atlanta, GA, USA, 2008, pp. 567-570.

[5] R. J. Baken, Clinical Measurement of Speech and Voice: An Introduction. College-Hill Press, 1987.