# On the use of ultrasound in speech communication systems

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

[1]Philips Research Laboratories Eindhoven, Netherlands
[2]Technical University of Eindhoven, Netherlands
firstname.lastname@philips.com

**Index Terms**: Ultrasound, Doppler Effect, articulatory movement, speech communication

## 1. Background and motivation

Traditionally, speech communication systems employ one or multiple air-conduction (AC) microphones to capture and process speech signals. These, however, are highly susceptible to background noise and reverberation. The performance of the corresponding systems therefore strongly depends on the acoustic characteristics of the environment in which they are being used. While numerous approaches to deal these problems have been investigated in the past [1, 2], a general solution does not exist and most solutions only work within certain constraints. This is especially the case for challenging environments - low signal-to-noise ratio (SNR), highly non-stationary noise or high reverberation. For this reason, researchers have started looking into sensors which are more robust to noise in so-called multi-modal speech communication systems. Such systems combine AC microphones with sensors like ultrasound, bone-conduction microphones, video cameras or electromyographs. In this work we use ultrasound as the additional modality. An advantage of ultrasound is that it is non-intrusive as opposed to bone-conduction microphones and electromyographs [3–6], since no skin contact is required. Also, ultrasound does not raise any privacy concerns unlike video [7, 8].

In the past, ultrasound sensors have been used successfully for voice activity detection [9, 10] as well as speaker and speech recognition [11–13]. Ultrasound-based speech communication systems are mostly set up so that the user is seated facing the sensor, which consists of an ultrasound receiver and an ultrasound transmitter. The transmitter emits an ultrasound signal which is reflected off the user's face and recorded again by the receiver. This reflection contains information about the articulatory movements during speech production.

Previously, ultrasound sensors in multi-modal speech communication systems were evaluated under the assumption that only articulatory movements are captured. In real-world applications, however, this assumption is unrealistic. In this work we investigate this and other practical considerations relevant for ensuring high performance of an ultrasound-based speech communication system. This work builds up on the methods described in [14].

## 2. Ultrasound Doppler processing in realistic scenarios

Most ultrasound-based speech communication systems use Doppler sensing to capture articulatory information. Hereby, an ultrasound beam or carrier signal is emitted by a transmitter and its reflections off a moving surface are captured and analyzed by a receiver. These reflections can contain one or more frequency components outside of the emitted signal due to the Doppler effect, which describes the change in frequency of a sound wave after being reflected off a moving object. The resulting frequency, $f_\mathrm{r}$, is given by

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

where $f_\mathrm{c}$, $v$, $c$ and $\Delta f$ are 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. Here, the moving surfaces of interest are moving articulators. An example of an ideal scenario where only articulatory movements are captured is shown in Fig. 1 for a 40 kHz carrier.

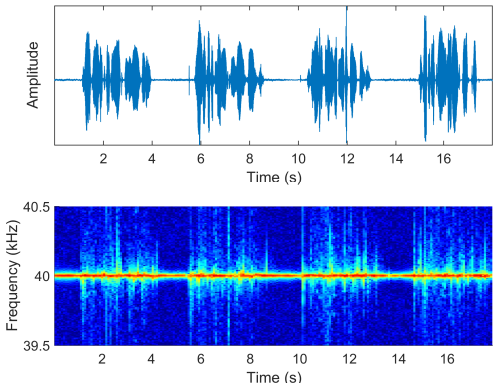


Figure 1: *Four utterances and the corresponding spectrogram of the ultrasound reflection. Articulatory information is encoded in the Doppler-shifted contributions around the carrier.*

In practice, however, the ultrasound sensor will capture movements of any object illuminated by its beam, which can lead to non-articulatory movement artifacts in the reflection. In this contribution, we start with the reflection model presented in [14], which allows us to detect and compensate for these artifacts. Furthermore, we design and evaluate articulatory features that can be extracted from the ultrasound reflection. We investigate their robustness by comparing features resulting from utterances which require similar articulatory movements, e.g., the diphones /pa/ and /ba/. Additionally, we investigate changes in articulatory features of a speaker for different scenarios which can lead to changes in articulation, e.g., Lombard speech.

Finally, we investigate a new use case for ultrasound Doppler processing by employing extracted articulatory information directly in a speech enhancement system and discuss the potential of this approach.

# 3. References


[1] P. C. Loizou, *Speech enhancement: Theory and Practice*. CRC Press, 2007.

[2] M. Brandstein and D. Ward, Eds., *Microphone arrays: Signal processing techniques and applications*, ser. Digital signal processing. Springer, 2001.

[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. Srinivasan and P. Kechichian, "Robustness analysis of speech enhancement using a bone conduction microphone - preliminary results," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

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

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270–287, 2010.

[7] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000, pp. 1–10.

[8] J. L. Crowley, J. Coutaz, and F. Berard, "Things that see," *Communications of the ACM*, vol. 43, no. 3, pp. 54–64, 2000.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] N. Cvijanovic, P. Kechichian, K. Janse, and A. Kohlrausch, "Improving the robustness of ultrasound-based sensor systems for speech communication," in *Proc. EUSIPCO*, 2015, pp. 889–893.