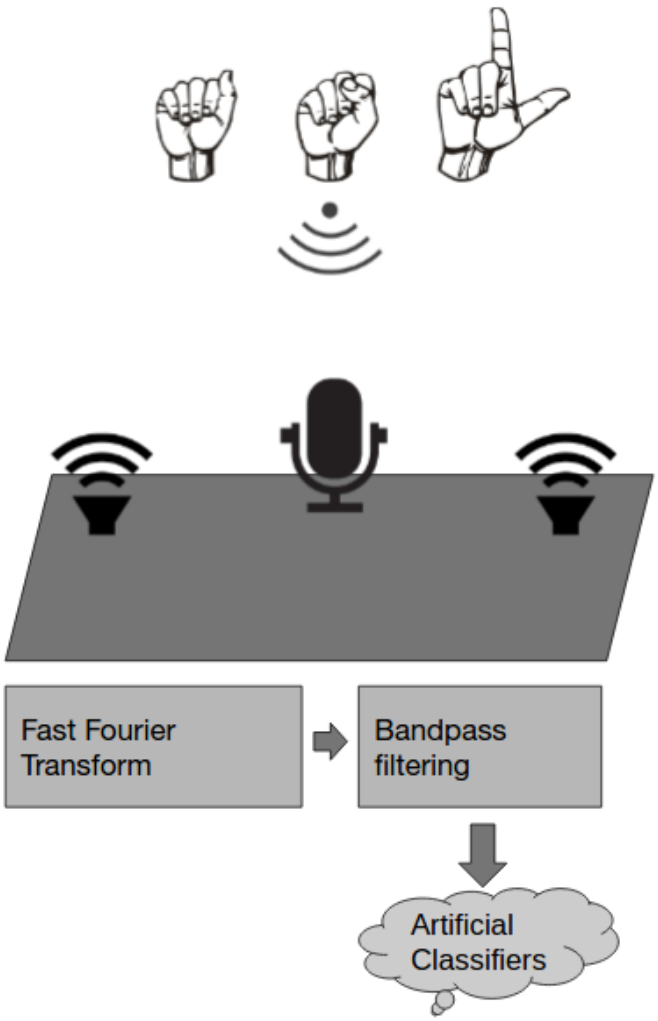
DopplerML, On Learning Ultrasonic Gesture Recognition

Matthew Harding (mharding), Tom Eliot (tke)



Abstract

Human-computer Interaction (HCI), the science behind major user-centered technologies, provides an interesting avenue for pursuing novel practical applications of machine learning. This work explores user-gesture learning using ubiquitous technology found in mobile computing platforms. Specifically, we resume previous work done that used the Doppler phenomenon to detect in-air user movement through simple spectral analysis of monophonic ultrasonic signals monitored via laptop microphone. For the purpose of classifying more complicated gesture examples, we extend the original hardware configuration to extract more spatial information from the Doppler interaction with the listening device. Using raw Fourier transform features, we observe that non-linear regression or ensemble classifier techniques preliminarily perform better for both longitudinal and lateral gesture recognition. Among our major contributions, we detail the potential feature-space to study next and make initial conclusions about the viability of machine learning for this area of HCI.

Motivation

The area of gesture recognition plays a central role in the science of Human-Computer Interaction, centered on creating more intuitive user interfaces for computing. Computers are projected to become increasingly pervasive and many current approaches to interface technologies, including gesture recognition, rely on compute-intensive computer vision systems for good performance.

Inspired by Microsoft Research's Soundwave project from 2012, we set out to attempt to bring the power of machine learning techniques to the unsown area of ultrasonic-based gesture recognition research. In their work, Sidhant Gupta, et al. present an informal analysis of the Doppler Effect for free space gesture recognition with the common laptop in any human environment. Their method was to continuously output a fixed ultrasonic tone from a monophonic speaker and Doppler shifts in this tone, with a frequency-domain peak offset larger than some threshold, detected by the microphone. These filtered shifts then allowed for classification as either approaching or receding user motions of differing velocities.

We began by hypothesizing that a stereophonic speaker system may provide more spatial information about a gesture if the two speakers output well-separated ultrasonic tones in the frequency-domain. We believe this extension to be necessary for machine learning to provide an improvement over the previous smart thresholding technique with the added encoding of lateral motion, as opposed to simply longitudinal. Our excitement also stems from the novel application of machine learning techniques to this area, so new that there is not yet a formal large public dataset available. The potential applications go beyond simple interfacial gestural play; we can imagine a more thorough Doppler-based prototype can roughly translate a signed language as a proof-of-concept.

Method

We utilize an approach to gestural recognition that requires minimal hardware. There exists an ultrasonic band of frequencies from 19kHz to 22kHz that a laptop or cellphone is capable of processing. To recognize gestures, techniques used commonly in radar and sonar may be applied, such as measuring Doppler shift and measuring time delay of reflections. Our hardware included a Macbook Air to emit sinusoidal tones of 20kHz and 20.5kHz in stereo from its speakers. When the user's hand moves with some relative velocity, Δv,in relation to the source of the impinging sound waves (which is stationary relative to the microphone), the waves are reflected and shifted in frequency according to the Doppler effect (1). The reflected signal is received using the integrated microphone.

$$\Delta f = \frac{\Delta v}{c} f_0 \tag{1}$$

To determine the frequency content of the reflected signal, we perform an FFT transformation on the received signal in real time using Hann windowing. We select a portion of the spectrum between 19.5kHz and 21kHz with the two sinusoidal tones centered. The result is a raw feature vector with 255 FFT bins across this frequency range, roughly representing the amplitudes of 500 frequencies. The machine learning gesture recognition software was implemented in the dataflow language, PureData, using the ml-lib [1] library. PureData provides many advantages when working with sound and signal data, including rapid prototyping and visualization, and powerful concurrency for real-time analysis. We looked at the performance of Support Vector Machine and K-nearest neighbors (KNN) regression techniques on a variety of gestures. Below, we've selected to reproduce only the KNN regression results.

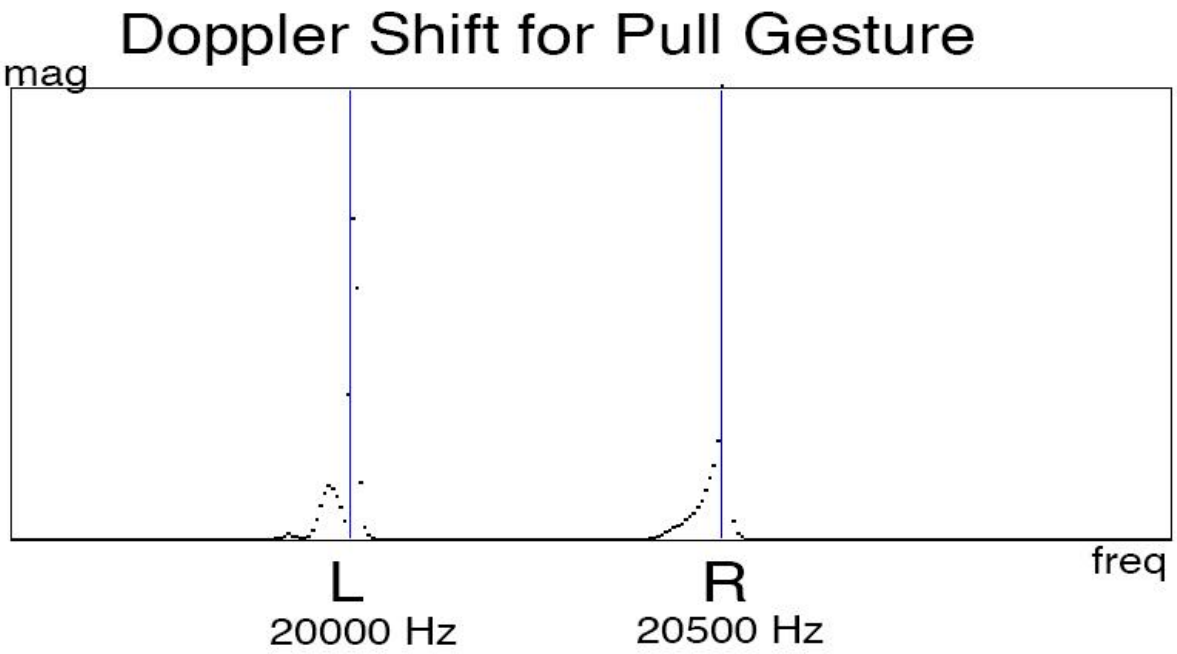


Fig. 1: Observed Doppler shift in ultrasonic bi-tonal for a longitudinal “pull” user-motion

Results

A K-nearest neighbors implementation was trained with 15 samples per class, and tested with 15 new samples each.

KNN

3 Classes

Gesture - Longitudinal dir. relative to microphone	Classification Accuracy
Still	100%
Push	100%
Pull	100%

3 Classes

Gesture - Lateral direction relative to microphone	Classification Accuracy
Still	100%
Swipe Left	85%
Swipe Right	80%

5 Classes

Gesture - Both Longitudinal and Lateral	Classification Accuracy
Still	33%
Push	33%
Pull	47%
Swipe Left	53%
Swipe Right	60%

Fig. 2: K-nearest neighbors classification performance for small datasets

For three gestures, the KNN classifier is effective. When qualitatively comparing a stereophonic machine learning approach with the thresholding technique in prior work [2] (a Doppler-controlled web application), we view this result is an improvement. The previous monophonic approach could only weakly detect left and right translational movements. Additionally, their thresholding technique is unable to differentiate between faster and slower push gestures, thus making it difficult to create a singular 'push' (including the retraction to the neutral position) gesture without misclassifying the following 'pull' gesture when returning the hand to a neutral position. As, perhaps, expected, KNN classification with 5 gestures was much less effective. We attribute the classification error to a lack of distinction between training classes (a swipe action contains "approach" and "recede" sub-motions relative to each speaker, which can confuse a classifier with little data) that stems from significant time-dependencies. More simply, the gestures were taught to the KNN classifier through noisy snapshots of the zoomed-in profile of the gestures' frequency-domains, and more complicated gesture recognition must also take into account the time-dependence of gestures for greater performance.

Ml-lib's functionality was such that we could easily drag-and-drop new supervised classification algorithms during our preliminary rounds of testing. Our best 3-class and 5-class performance was observed using K-nearest neighbors as opposed to linear classifiers like a Support Vector Machine. We believe this non-linearity to be due to the relationship of the observed Doppler effect with gesture speed, as well as significant sample noise. This is not to say that a simple non-linear classifier may demonstrate the best performance for this small dataset, as the 5-class performance doesn't show much promise. In fact, the smarter approach may have been to construct an ensemble classifier that would first distinguish between motion types (lateral or longitudinal) and second distinguish between motion direction (receding and approaching microphone), and take advantage of the perfect classification rate for the two 3-class sample sets.

Future Work

Working in a dataflow language like PureData provided advantages in streamlining our workflow. However, the ml-lib library constrained our maximum feature vector to length 255, and our limited experience in the PureData developing environment restricted our feature data to be raw snapshots of the fourier domain. We know that there is much higher resolution data that may be collected on the frequency spectrum and trained upon. Additionally, more complex features, such as spectrograms, additionally encode time information that was lacking our in

sample set. A spectrogram, or, essentially, the fourier domain of a signal sampled over time, is a 2-dimensional feature that shows promise for gesture recognition. There is a precedent for this form of signal recognition in Daniel Nouri's work on the use of deep neural networks on sonogram data in the classification of animal noises e.g. whale and bird calls. In this work, he frames spectrogram classification as an image classification problem, and his technique attained roughly 97% accuracy at competition [3]. We would like to harness the same deep-learning technique for gesture recognition, as we believe there are minute differences in the signals over time that can help distinguish complex gestures and provide greater gesture speed-invariance. We would also like to try to apply the technique of sonar to learned gesture recognition, where the reflection delay is measured as opposed to the Doppler shift.

Conclusions

Our exploration shows early results in applying Doppler-phenomenon feature classification in the ultrasonic range for practical use in gesture classifying applications. Our innovations are two. First, the application of AI to ultrasonic gesture recognition. Second, the use of stereo in ultrasonic gesture recognition. We hope that the accessibility of the hardware required for this technology will motivate further exploration of this technique. We believe that classification performance and number of gestures can be improved significantly with continued work. The use of existing hardware has exciting implications for the application of the technology in HCI and gaming.

Our PureData implementation is available here:
https://www.dropbox.com/s/f55pj1zd7fsinnd/touch-and-activate-stereo.pd

It must be used with ml-lib [1].

References and Related Work

[1] https://github.com/cmuartfab/ml-lib
[2] https://github.com/DanielRapp/doppler
[3] http://danielnouri.org/notes/2014/01/10/using-deep-learning-to-listen-for-whales/

SoundWave: Using the Doppler Effect to Sense Gestures
Sidhant Gupta, Dan Morris, Shwetak Patel, Desney Tan
Proceedings of ACM CHI 2012, May 2012

Touch & activate: adding interactivity to existing objects using active acoustic sensing
Makoto Ono, Buntarou Shizuki, and Jiro Tanaka.
In Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST '13)