Sumit Basu and Alex Pentland, "Smart Headphones," Proceedings of CHI 2001, Seattle, WA

Smart Headphones

Sumit Basu
MIT Media Laboratory
E15-383, 20 Ames Street
Cambridge, MA 02139 USA
+1 617 253 0378
sbasu@media.mit.edu

Alex Pentland
MIT Media Laboratory
E15-383, 20 Ames Street
Cambridge, MA 02139 USA
+1 617 253 0648
sandy@media.mit.edu

ABSTRACT
In this paper, we describe the "smart headphones" system, which detects and relays speech sounds in the environment through the user's headphones. This allows the user to be aware of speech events and remain involved in conversations while using headphones.

Keywords
headphones, auditory awareness, speech processing

INTRODUCTION
Though our ears are wonderful instruments, there are times when they simply cannot handle everything we need them to. This can be for a number of reasons - physical boundaries (headphones or walls), conflicting sound sources, high cognitive loads, and of course hearing loss. While there are many sounds we can miss due to these reasons, among the most critical are the sounds of speech and conversation because they are so critical to our social interactions. The goal is to detect these types of auditory events and to selectively make the user aware of them.

In order to achieve this, we rely on two pieces of our recent work in signal processing. The first is a speech-detection algorithm, which is able to robustly discriminate the sounds of speech from other sounds. This work is described in a forthcoming paper and we give a very brief summary of it here. The second is a method to determine where the speech sounds are coming from using multiple microphones on a body-based array. We developed this technique in [1] for wearable computers. The user interface allows the user to set the directional sensitivity for speech – she or he can have speech from only particular directions come through.

The application scenarios for this are many: our personal favorite is that of "smart headphones," which allow a user to listen to her or his favorite music at top volume without disturbing others or losing awareness of the conversational scene. When speech starts coming in, s/he can either have it played through the headphones as s/he continues listening to the music or stop the music to pay full attention to the conversation. In either case, it alleviates the annoying but common situation of having to come up to the headphone-wearer, tap her or him on the shoulder, and wait for her or him to take off the headphones before speaking to him.

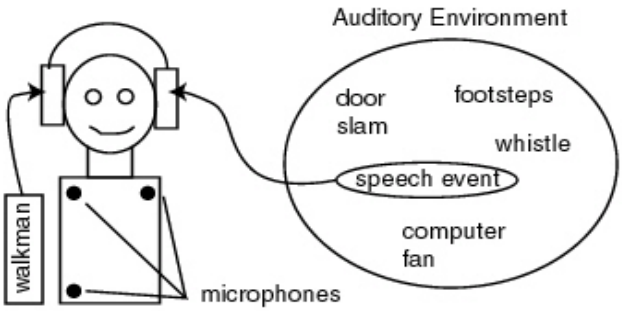

Figure 1: The smart headphones system. The system detects speech sounds in the environment and passes them through to the user's headphones.

The user can also selectively listen for/amplify speech only from her or his left, for instance, where a friend may be sitting in a crowded plane, allowing her or him to disregard speech from others. Another "smart headphones" scenario is for people working in high-noise environments - airport runways, steel foundries, etc., where hearing protection is a necessity. The proposed system could allow such workers to have normal conversations without taking off their protective gear. Hearing aid applications are in a similar vein, though here the slight delay introduced by the detection algorithm could hurt speechreading performance. Moving away from headphones, we see a variety of other applications as well. For instance, the microphones could be placed outside the user's office door, set off to relay audio only when speech is coming from directly in front of the door. The speech of people passing through the hall would be ignored, while a visitor's speech would come clearly through on the user's speakers. In this paper, we first discuss the methods we use to implement our system. We then discuss some preliminary results from using this system and close with a proposal for future work.

SPEECH PROCESSING METHODS

Speech Detection
The task of speech detection in open environments (i.e., without using a headset microphone) has received increasing attention in the past few years due to the growing interest in speech input for cars, homes, and so on.

Most of these methods have relied on the "voiced" segments of speech (vowels) occurring in certain frequency bands (as in the recent work of Wu and Lin [2]). Our method has gone a step further, looking at the harmonic structure of the frequency content in these voiced segments. As a result, our algorithm cannot be fooled by other sounds that simply fall into the right frequency range – only sounds with a "voiced" quality (such as musical instruments) will be mistakenly detected as speech. We have yet to perform a formal evaluation of this algorithm, but in initial experiments it caught 84% of utterances in their entirety and at least some part of 91%, with less than 10% false positives (other sounds mislabeled as speech).

Estimating Speaker Direction

The second signal processing method, that of seeking the direction where sound is coming from, is an old one – it is known as "direction of arrival estimation" and has been well studied by the radar and sonar communities [3]. However, it is somewhat more difficult to apply these methods to speech, especially with microphones that may be moving about on the body. We describe how we encounter these issues in [1]. With three microphones on the body placed as shown in figure 1, we are able to determine the direction of the sound both horizontally and vertically with respect to the user's body. With the system as described in our earlier work, we were able to estimate the direction speech was coming from with 30 degrees about 88% of the time.

Combining the Methods

We combined these two algorithms in a serial manner: first running the speech detector and then the direction estimation algorithm. Only audio that was marked by the speech detector as containing speech was considered by second algorithm. At this point, the system would have chunks of speech with labels of which direction they were coming from. Based on the user's direction selection, the appropriate chunks would be mixed in with the other audio going to the user's headphones.

Practical Considerations

All of our speech processing methods work in real time on a Pentium II/III processor and consume only a small fraction of the CPU. However, because the speech detection algorithm needs to see a certain amount of the incoming signal before it can decide whether it is speech, there is an inherent delay to the system. In our experience, this has to be a minimum of ¼ second to allow reasonable performance of the speech detector. The resulting delay is not signficant in terms of conversational dynamics, since expressions and gestures tend to be on a slower timescale than this [3]. However, as we found in our prelminary user tests, the effect of listening to one's own voice with a slight delay is very irritating, and in fact makes it difficult to speak. As a result, we use the information from the speaker direction algorithm to automatically filter out speech coming from the speaker's mouth. The disadvantage of this is that the user hears her or his voice muffled through the headphones/music and tends to speak more loudly than necessary to compensate. We plan to compensate for this in a future version by passing sound coming from the direction of the user's mouth directly into the headphones without first running the speech detection algorithm. Since the only sound sources in this direction are the speaker's mouth and sources above his head, this should not result in many extraneous sounds being passed through as speech.

PRELIMINARY RESULTS

While we have not yet done extensive user testing on our system, we have tried it on four subjects. Two of these subjects were not told what the system was supposed to do or how it worked. The users were given two controls: speech loudness, sensing direction, and beam width of sensing direction.

The reactions of the naive users was the most impressive – they were given the headphones with music playing, and assumed they were being asked to listen to a particular song. As soon as someone in the room spoke, they became immediately aware of what the system did. The setting of the sensing direction control had to be explained, but was readily understood by both subjects.

All of the subjects felt that using the system let them be more aware of their environment with their headphones on, and felt less excluded when a conversation would start. The main complaints was that they would sometimes miss small chunks of speech, or that the style of music they were listening to made it impossible to understand the utterance when it was mixed in.

FUTURE WORK

There are a number of ways in which we wish to extend and improve this system. First, we want to experiment with reducing the volume of other audio in the headphones when speech is being played. Second, we want to give the user more control over the sensitivity of the speech detection algorithm. Third, we would like to give the user the option of not having speech passed through immediately, but instead hearing a "call waiting" tone that signified speech had occurred, then storing the speech for them to browse at their convenience.

ACKNOWLEDGMENTS

We wish to thank Brian Clarkson for his instrumental role in helping develop the speech detection algorithm.

REFERENCES

1. Basu, S., Schwartz, S. and Pentland, A. Wearable Phased Arrays for Sound Localization and Enhancement. In *Proceedings of the IEEE Int'l Symp. on Wearable Computing* (Atlanta GA, October 2000), IEEE Press, pp.103-110.
2. Wu, G., and Lin, C. Word Boundary Detectio with Mel-Scale Frequency Bank in Noisy Environment. In *IEEE Trans. on Speech and Audio Processing*, vol. 8, No. 5, IEEE Press, 2000.
3. Bolinger, D. Intonation and Its Parts: Melody in Spoken English. Stanford University Press, 1986.