Sumit Basu, Brian Clarkson, and Alex Pentland. "Smart Headphones: Enhancing Auditory Awareness Through Robust Speech Detection and Source Localization." Appears in Proceedings of the Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP '01). Salt Lake City, UT. May, 2001.

SMART HEADPHONES: ENHANCING AUDITORY AWARENESS THROUGH ROBUST SPEECH DETECTION AND SOURCE LOCALIZATION

Sumit Basu, Brian Clarkson, and Alex Pentland

MIT Media Laboratory, Cambridge, MA 02139

ABSTRACT

We describe a method for enhancing auditory awareness by selectively passing speech sounds in the environment to the user. We develop a robust far-field speech detection algorithm for noisy environments and a source localization algorithm for flexible arrays. We then combine these methods to give a user control over the spatial regions from which speech will be passed through. Using this technique, we have implemented a "smart headphones" system in which a user can be listening to music over headphones and hear speech from specified directions mixed in. We show our preliminary results on the algorithms and describe initial user feedback about the system.

1. INTRODUCTION

There are many situations in which our ears are not sufficient to analyze the auditory scene around us. 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 the general problem of auditory scene analysis [1] has received a great deal of attention, it is extremely difficult to deal with all possible forms of environmental sound. We focus here on a much smaller domain, namely speech and conversations, because of their critical role in our social interactions. The goal is to detect these types of auditory events and to selectively make the user aware of them.

Our approach combines speech detection and speech source localization techniques. We have developed a robust speech detection algorithm for far-field microphones (i.e., does not a require close-talking/noise canceling microphone) that detects voiced sounds and combines them when they co-occur into utterances. This mechanism is fairly robust but is sensitive to certain kinds of harmonic sources/noise. The second stage is a source localization algorithm that determines the most likely direction of utterances from a flexible array of three or more microphones. The user interface allows the user to set the directional sensitivity for speech – 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 his favorite music at top volume without disturbing others or losing awareness of the conversational scene around him. When speech starts coming in, he can either have it played to him through his headphones as he continues listening to his music or stop the music to pay full attention to the conversation. In either case, it alleviates the annoying and familiar situation of having to come up to the headphone-wearer, tap him on the shoulder, and wait for him to take off the headphones before speaking to him. The user can also selectively listen for/amplify speech only from his left, for instance, where his friend may be sitting in a crowded plane, allowing 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 describe our techniques for speech detection and speech source localization. We then show some preliminary results from using this system and close with a discussion of our future work.

2. BACKGROUND AND METHODS

There has been a large body of prior work in both speech detection and source localization. In the interests of brevity, we will only touch upon the most relevant work and describe how our work relates to it.

2.1. Speech Detection

As interest in speech applications for open environments has grown, speech detection in noisy environments has received increased attention from speech researchers. For instance, there is the work of Junqua et al. [2] which presents a number of adaptive energy-based techniques, the work of Huang and Yang [3] which uses a spectral entropy measure to pick out voiced regions, and most recently the work of Wu and Lin [4], which extends the work of Junqua et. al by looking at multiple bands and using a neural network to learn the appropriate thresholds. The basic approach of these methods is to find features that allow detection of voiced segments (i.e., vowels) and then group them together into utterances. We found this compelling, but noted that many of the features suggested by the authors above could be easily fooled by environmental noises.

As a result, we sought a feature that was more specific to speech than particular bands of spectral energy. We found this in the

Samsung v. Techiya
IPR2024-01004
Exhibit 1028

banded structure of voiced segments - as we know from the source-filter model of spech, the Fourier transform of the glottal pulse convolved with the vocal tract is the vocal tract transfer function *multiplied* by a sequence of peaks at integer multiples of the pulse frequency. Furthermore, since the pitch varies continuously within voiced segments, these lines are continuous as well. This results in a clear set of striated lines through every vowel, as shown in figure 1 below. When such lines exist in a harmonic relation for a long enough time, we can be quite certain that voicing is present. Such a feature will be robust to most environmental noises, but will of course be susceptible to certain musical instruments such as tubas, and other sources which have a similar source-filter model.

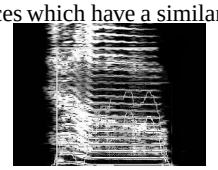

Figure 1: Continuous, striated lines running through a voiced segment

In order to make use of this feature, we first go through several preprocessing steps. In order to catch the banded nature of the vowels, we use a 1024 point FFT at a framestep of 64 samples (sampling at 16 kHz). After computing the log of the power spectrum, we compute a running expectation of each band $X_k[t]$ in $X_m, k[t]$with a simple IIR filter in time:

$$X_{m,k}[t] = (1-\alpha)X_{m,k}[t-1] + \alpha X_k[t] \qquad (1)$$

The coefficient α begins at 10^{-2}, allowing the system to adapt rapidly at first, and is eventually decayed to 10^{-5}, allowing for slow but continual adaptation (note that if frames are marked as containing voicing, they are not included in the adaptation). We do the same procedure for the square of each band, $X_k^2[t]$. We can then normalize each band by its running mean and variance:

$$X_{n,k}[t] = \frac{(X_k[t-1] - X_{m,k}[t])}{\left(X_{m,k}^2[t] - (X_{m,k}[t])^2\right)^{1/2}} \qquad (2)$$

We do the computation above with one caveat: if the standard deviation of $X_k[t]$ (i.e., the denominator above) is less than one, we do not divide by it. This prevents low-energy bands with a small variance from producing large spikes from small amounts of energy.

The next task is to find all of the candidate peaks in the normalized log power spectrum. In order to prevent small ripples at either high or low energies to be counted as peaks, we use a simple hysteresis mechanism. There are two thresholds: p_{min}, which is the minimum value the power must reach in order to be considered a peak, and v_{max} which is the value to which the power must go down within the following valley before another peak can be considered to begin. Once the beginning and ends of peak periods are found, the maximum values within each period are chosen as the actual peak locations. This process is illustrated in the figure below (figure 2).

With the peak candidates found, we would like to determine which of them fall into a harmonic relationship. However, we found that trying to detect bandedness on a per-frame basis was unreliable due to the noise in the signal (as also noticed by Wu et al. in [5]). Instead of summing adjacent frames, though, which is equivalent to taking longer framesizes, we took the approach of tracking the lines over time and then checking for a harmonic relation. The tracking algorithm is simple – every time a peak is found, it is checked whether it could be a continuation of any of the existing lines. To be such a continuation, it must be within two samples (in the FFT domain) of the previous location. If it is not, a new line is started. In addition, each line is allowed to have gaps of up to two frames to help deal with corruption by noise. This process results in many spurious "line pieces." The number of candidate lines can be greatly reduced by requiring candidate voicing lines to be of a minimum length l_{min} – in our case, 20 frames.

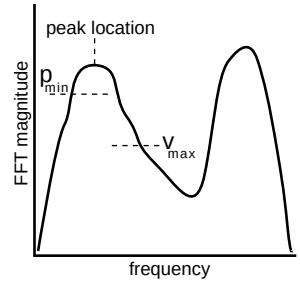

Figure 2: Peak thresholds p_{min} and v_{max}

Once the sufficiently long lines have been marked out, we check to see if they form a harmonic relationship. Note that in order to know whether the lines were long lines, we must wait the same number of frames as the minimum length for a voicing line. Looking back 20 frames in time, then, we go through pitch candidates from 30 Hz to 350 Hz at a stepping of about 3 Hz. To do this, we go through all multiples of the candidate pitch, f_{cand}, up to f_{max} (8 kHz), incrementing the number of bands k for that candidate if the nearest long line's frequency (f_l) is within 2 samples. This calculation can be written as:

$$k = \max_{f_{cand}} \sum_{n=1}^{\lfloor f_{max}/n \rfloor} (|n f_{cand} - f_l| < 2) \qquad (3)$$

Once the criterion for bandedness ($k > 5$, in our experiments) has been achieved, we find the longest line that fits this banded relation and trace it from its beginning to its end, and mark that region as being voiced.

The next step is to group voiced segments into utterances, so that we catch the unvoiced parts of the speech as well. At this point, we are doing this using only proximity in time – if another voiced segment occurs before the maximum within-utterance silence gap (t_g) has passed, it is considered part of the utterance. As a result, we cannot tell if an utterance has ended until t_g frames have passed. Since t_g is typically on the order of a second, this can be too long a gap for some applications. We can counteract this by either reducing t_g or making the best possible decision at a sooner time. The latter policy results in more false positives (we will sometimes say we are in an utterance when we are actually not), but will not cause us to miss speech and let us work with the much shorter latency of the voicing decision (l_{min}). The figure below illustrates the detection of voiced segments and utterances (figure 3).

The main weakness of this algorithm is the large number of thresholds/parameters. Fortunately, they seem to be fairly robust to

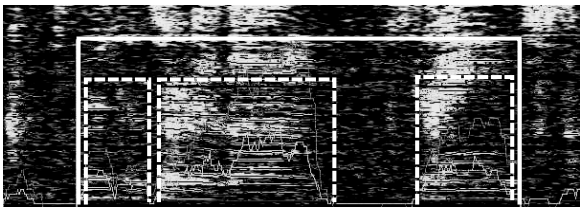

Figure 3: Detected voiced (small dashed boxes) and utterance (large solid box) segments. The contrast has been normalized to make the segmentation more clear.

various microphones and large environmental variations (as shown in the results section below). This is due mostly to the adaptive pre-processing step. It would be quite possible to train most or all of these parameters, and we are pursuing methods to do this at the moment.

2.2. Source Localization

There is a long history of work in source localization in the radar/antenna community, and we cannot begin to give a complete history of it here. However, there are three assumptions typically made in this work that are not satisfied in our case. First, the signals dealt with in that community are either of a known form (a pulse that is sent out and modified in parametric ways upon reflection) or within a small frequency range. Second, the signals are coming from a great distance, allowing for a "plane wave" assumption (i.e., the wavefronts appear as straight lines instead of circles because the source is so far away). Third, the array geometry is typically assumed to be both known and fixed. In our case, we have to deal with an unknown signal (usually speech) that has a very wide spectral range and is often coming from nearby (the user's mouth!). To make the problem even worse, in the cases where the microphones are worn on the body (typical for the smart headphones application), our array geometry is unknown and constantly in motion.

Because speech does not satisfy the usual assumptions, phased arrays have not been widely used for speech processing. There are a number of works that do beamforming (signal enhancement) with speech [6, 7, 8, 9], including [8], who develop an array built into a pair of eyeglasses that do fixed beamforming for a hearing aid. The use of arrays for speech localization has been even more limited – [7] and a few others. Last, as far as we know, there is no other work before ours on building arrays on the body for source localization.

We give a complete account of building how we do localization from a body-based flexible array in [10], but we summarize some key details here. Our method for delay estimation between the microphones is straightforward – one simply has to be careful about the spacing and frequency constraints (see [10]). The interesting piece is estimating the actual incoming direction of the sound. Since we cannot make a far-field assumption, the constraint of constant delay between two fixed points yields one side of a hyperbola (the other side is not relevant since sound only propagates in a positive direction). In 3D, this is a hyperboloid (the hyperbola is rotated around the axis connecting the two mics. For three microphones, the intersections of the hyperboloids form parabolas in space, typically going through the user's body. Since the body acts as an acoustic shield, only one side of the parabola is relevant. The proximity of the sources prevents us from using the asymptotes of these parabolas, and furthermore it is impossible for us to solve explicitly for the source direction since the user may reconfigure the array every time he puts it on, putting the microphones on different parts of the clothing.

As a result, we choose to learn a mapping from the delays to a source direction. At this stage, we are using the simplest possible model – an affine mapping between delay space and 2D orientation, i.e.,

$$\begin{bmatrix} \theta & \phi \end{bmatrix}^T = A \begin{bmatrix} d_1 & d_2 & 1 \end{bmatrix}^T \tag{4}$$

At this point, the user steps through a sequence where he snaps to his front, right, left, below his waist, and then speaks a short utterance (all upon cue from the program). We now have four labeled 2D $[d_1 d_2]$ pairs, $[d_1^1 d_2^1]$ through $[d_1^4 d_2^4]$. Writing the i, j element of A as a_{ij}, we can rewrite equation 4 as

$$\begin{bmatrix} d_1^1 & d_2^1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1^1 & d_2^1 & 1 \\ d_1^2 & d_2^2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1^2 & d_2^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \end{bmatrix} = \begin{bmatrix} \theta^1 \\ \phi^1 \\ \theta^2 \\ \phi^2 \\ \vdots \end{bmatrix} \tag{5}$$

If we write this equation as $Da = b$, we can easily find the least-squares solution for a as:

$$\hat{a} = (D^T D)^{-1} D^T b \tag{6}$$

The results of the phase estimation and thus the mapped source location are somewhat noisy, so we use a simple dynamic programming algorithm to enforce stability on the location values by imposing a cost on location transitions. An example sequence of tracking 7 speaker changes in a 10 second period is shown in figure 4 below.

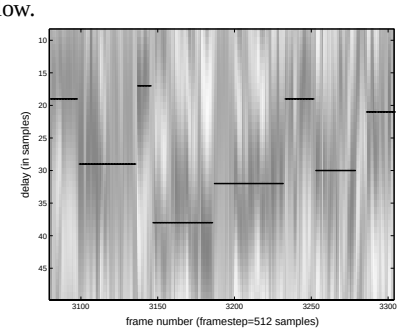

Figure 4: Speaker change detections using the dynamic programming algorithm overlaid on the raw cross-correlogram. Darker values on the cross-correlogram signify higher correlation values. The contiguous line segments correspond to different speakers. Ten seconds of speech are shown; all seven speaker changes are caught by the algorithm, including a 1/3 second interjection.

2.3. Combining the Methods

The speech detection and source localization are combined serially – the source localization algorithm is activated only on the segments detected as containing speech. The user is able to specify

what range of locations he wishes to accept speech from and independently adjust the volume of the speech. Audio is then passed through to the user only when speech is detected within the appropriate spatial range. In order to keep the latency short, we play frames that are classified as being within an utterance but without waiting for the silence gap of t_g frames to have passed. As we mentioned earlier, this allows to get all the speech frames at a low latency at the expense of playing some extra frames at the end of each utterance.

3. HARDWARE SETUP AND PROCESSOR LOAD

We tried two different hardware setups for this system. The first one involved headphones plugged into a desktop computer with microphones suspended overhead. The second had smaller microphones place on the body (two on the chest, one near the waist). On the desktop machine (PIII 700 dual processor), the algorithms above took from 5% to 10% of the machine's CPU, while on the mini-laptop (PI 200 MHz), they took from 10% to 30%. The loads were this low due to heavy use of Intel's SIMD instruction sets through the Intel Performance Libraries.

4. PRELIMINARY RESULTS

At this point, we have not tested our speech detection method on a large database, but from small experiments in office and open lab space environments the detector catches 82% of utterances "whole," i.e., the entirety of the utterance is marked out, and 91% of utterances are partially marked. 12% of the marked segments were false positives. The source localization estimated the correct direction within 30 degrees 88% of the time. We are currently doing more formal experiments for both of these components in a number of environments/noise conditions, and will report on these results in a later paper.

We have tested the system on four different users in a noisy office environment (doors opening/closing, fan noise, books dropping on tables, etc.), two of whom had no prior knowledge of the system or what it was supposed to do. All four reported that the system significantly improved their awareness of the speech around them without distracting them with all of the non-speech sounds, and found they could carry on a conversation with music playing. These tests led us to an interesting discovery – since there is a fixed lag for the speech detection algorithm, the user's own speech was echoed back to him with a slight delay. Users found this exceedingly annoying and found it difficult to speak with the feedback on. However, because of our source localization stage, it was easy to eliminate this effect by simply ignoring speech coming from the user's location. Since the user is at a fixed phase with respect to the microphones in the wearable setting, this works smoothly even while he is moving around. There is one caveat here – because the user's own voice is muffled by the headphones and the music, he will tend to speak much more loudly than necessary. A simple solution for this would be to pass sound coming from the user's direction through without running the speech detection. This would require running the source localization algorithm on all incoming data, but would make the user's interaction in the conversation much more natural.

5. CONCLUSIONS AND FUTURE WORK

We have demonstrated a mechanism for enhancing auditory awareness through a combination of speech detection and source localization techniques, and have successfully developed a "smart headphones" application in an office environment. Though our results are still preliminary, the methods we have described seem to perform robustly in moderately noisy environments. Furthermore, user feedback indicates that the system is indeed capable of allowing a headphone-wearer to be aware of the speech in the environment around him and to take part in a conversation without taking off his headphones.

There are a number of directions we wish to take this work. First, we want to make a more formal evaluation of all the components of this system as well as the system as a whole. Next, we wish to explore some of the other application areas we described in the introduction. Furthermore, we would like to see what other types of speech filtering may be useful in this scenario. For example, we could use speaker identification techniques to block out or pass through speech from particular individuals. Last, we would like to explore additional user interface mechanisms for getting the speech to the user. Instead of having all speech streamed to the user, for instance, the system could notify him of speech from a particular direction with a spatialized tone and allow him to browse through the last few utterances at his leisure.

6. REFERENCES

[1] Albert Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[2] Jean-Claude Junqua, Brian Mak, and Ben Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 406–412, 1994.

[3] Liang sheng Huang and Chung ho Yang, "A novel approach to robust speech endpoint detection in car environments," in *Proceedings of ICASSP'00*. 2000, pp. 1751–1754, IEEE Signal Processing Society.

[4] Gin-Der Wu and Chin-Teng Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 541–553, 2000.

[5] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, and X. Menendez-Paul, "A robust speech detection algorithm for speech activated hands-free applications," in *Proceedings of ICASSP'99*. 1999, pp. 2407–2410, IEEE Signal Processing Society.

[6] H. Cox, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[7] F. Khalil, J.P. Jullien, and A. Gilloire, "Microphone array for sound pickup in teleconference systems," *Journal of the Audio Engineering Society*, vol. 42, no. 9, pp. 691–699, 1994.

[8] R.W. Stadler and W.M. Rabinowitz, "On the potential of fixed arrays for hearing aids," *Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1332–1342, 1993.

[9] Sumit Basu, Michael Casey, William Gardner, Ali Azarbayejani, and Alex Pentland, ""Vision-Steered Audio for Interactive Environments"," in *Proceedings of IMAGE'COM*, Bordeaux, France, May 1996.

[10] Sumit Basu, Steve Schwartz, and Alex Pentland, "Wearable phased arrays for sound localization and enhancement," in *Proceedings of the IEEE Int'l Symp. on Wearable Comp. 2000*. 2000, pp. 103–110, IEEE Signal Computer Society.