Smart Headphones: Enhancing Auditory Awareness with Speech Detection and Source Localization

Sumit Basu, Brian Clarkson, and Alex Pentland
MIT Media Laboratory

SPEECH DETECTION

PRIOR WORK

- Recent interest in far-field speech has led to work on speech detection
- [Junqua et al. 94] use adaptive energy, [Huang and Yang 2000]use spectral entropy to pick out voiced regions, [Wu and Lin 2000] use energy in multiple bands

OUR METHOD

- We wanted a feature more specific to voiced sounds - make use of the harmonic structure of speech (figure 1) which results in banded spectral lines at multiples of the pitch (F0)
- We start by estimating spectral mean and variance in unvoiced regions and then normalizing:

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

- We then identify peaks in the spectrum using a simple hysteresis (fig. 2)
- Each peak is then followed in future frames to build up "spectral lines." If a line extends for a minimum number of frames, it is kept as a potential spectral line
- Looking back a fixed lag L in time, we compute the bandedness of frame t-L by iterating over all possible vocal pitches and counting the number of spectral lines that could be accounted for by voicing at such a pitch:

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

- If k exceeds a threshold, we follow the longest spectral line in the group from its beginning to its end and mark the duration as a voiced chunk
- Finally, voiced chunks that are within a threshold of each other in time are grouped together as part of an utterance. The window of an utterance is slightly extended at its onset to account for initial/final consonants (fig. 3)

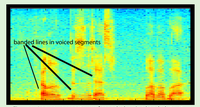


Figure 1: Log power spectrum of speech showing banded structure

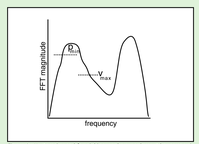


Figure 2: Hysteresis used for picking peaks: a peak must have magnitude at least pmin, but a new peak will not be detected until the level has dropped below vmax.

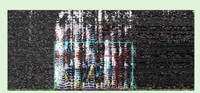


Figure 3: Results of the proposed speech detection algorithm. The red boxes outline detected voiced segments, while the light blue box shows the grouping of voiced segments into an utterance. The dark blue lines show the tracking of the spectral lines.

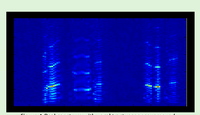


Figure 4: Real cepstrum with vocal tract response removed

RESULTS

- In office environments, the detector correctly labels 82% of utterances whole (the entirety of the utterance is marked as speech)
- 91% of utternances are partially marked

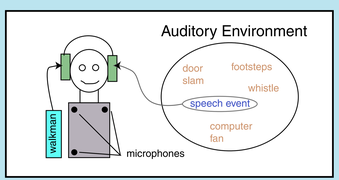
CURRENT WORK

- Probabilistic approach: create an HMM whose outputs describe the features of voiced/unvoiced states. States are labeled, so training is simple. Greatly reduces number of thresholds.
- Integrate the pitch estimation task using the cepstrum (fig. 4)
- Choose candidate peaks by using multiscale peak-finding

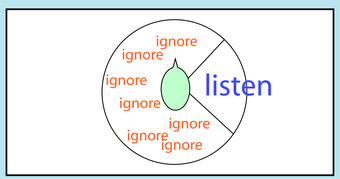
THE PROBLEM SCENARIO

Headphones are an excellent means to enjoy audio content without disturbing others, but they have the effect of isolating us from the social world. Here is a typical scenario where someone is trying to get information from the headphone wearer.

THE PROPOSED SOLUTION



- Smart Headphones: mix appropriate sounds from the environment into the headphone output (with a small delay)
- Perform speech detection and only pass speech sounds through to the user
- Use source localization to allow the user to select particular directions to attend to and ignore others



PROTOTYPE AND APPLICATIONS

PROTOTYPE

- Running on an intel pentium-class processor making use of the intel performance libraries. Overhead and wearable mics
- CPU usage: 3% for PIII-700, 50% for PI-166
- Tested on a small group of users, two of which were not told the intention of the device, but understood immediately upon using it. All reported enhanced awareness of environment.

APPLICATIONS

- Basic form: software application for a computer or part of a portable music player/headphone set
- Hearing protection: airport workers could have conversations without removing protective headsets
- Smart Intercom: for people knocking at office door - only pass through speech of people directly in front of door

SOURCE LOCALIZATION

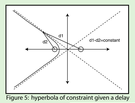


Figure 5: hyperbola of constraint given a delay

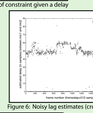


Figure 6: Noisy lag estimates (cross-correlation)

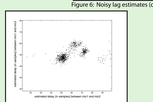


Figure 7: Lag estimates showing clusters

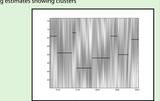


Figure 8: Decoding speaker changes with DP

PRIOR WORK

- Great deal of work exists on source localization, but relatively little on speech: [Khalil et al. 1994] for teleconferencing systems; no work on body-based arrays until [Basu et al. 2000]

OUR METHOD

- Basic cross-correlation based method - lag between microphones gives us a hyperbola of constraint (figure 5)
- Initial result of phase is noisy but results in distinct clusters (figure 6, 7)
- Use dynamic programming algorithm to decode optimal path (figure 8)
- Microphone geometry is flexible (microphones placed on body), so learn mapping from lag to direction with least-squares:

$$\begin{bmatrix} d_1^1 & d_0^1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1^1 & d_0^1 & 1 \\ d_1^2 & d_0^2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_1^2 & d_0^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \end{bmatrix} = \begin{bmatrix} \theta^1 \\ \phi^1 \\ \theta^2 \\ \phi^2 \\ \vdots \end{bmatrix}$$

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

RESULTS

- Correct direction within 30' 88% of the time (lower in highly reflective areas)

FUTURE WORK

REFINING ALGORITHMS

- Probabilistic pitch-tracking and voicing detection
- Train and test methods on larger database with a variety of microphones and recording conditions
- Try onset-based algorithms for source localization to eliminate problems with reflections
- Detect wearer's speech and pass through instantly to avoid irritating 1/4 second feedback

FILTER SELECTIVITY

- Use speaker-ID techniques to do person-specific blocking
- Using prosody to discriminate between conversations and queries directed at the user