Adaptation of Tandem HMMs for Non-Speech Audio Event Detection

Mark Hasegawa-Johnson, Xiaodan Zhuang, Xi Zhou, Camille Goudeseune, and Thomas Huang

These slides:
http://www.isle.uiuc.edu/slides/2009/Hasegawa-Johnson09ASA2.pdf

ASA Spring Meeting, May 21, 2009

ILLINOIS

Outline

Task Definitions

Task #1: Acoustic Event Detection

- Detect non-speech acoustic events (door slam, chair movement, paper shuffle) in a meeting room
- What happened when?

Task #2: Speech Phone Classification

- Given an acoustic spectrum x_i, specify the phone label y_i
- A heavily-studied problem, therefore the baselines are well understood

Task #1: Non-Speech Acoustic Event Detection

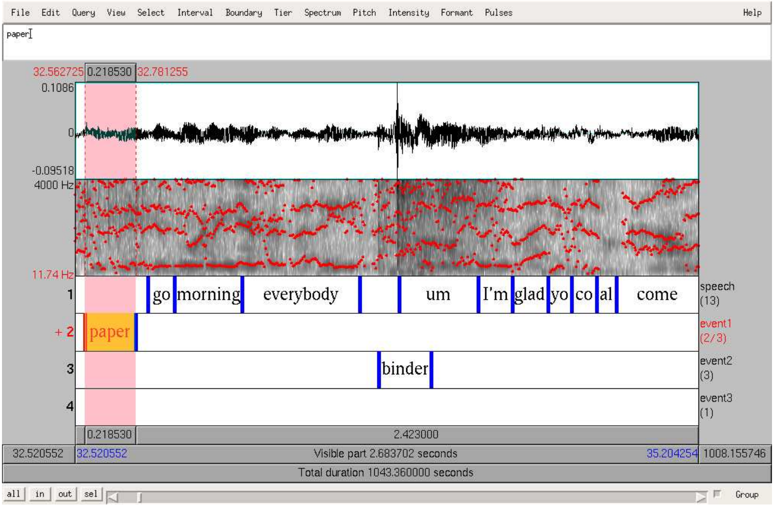
Motivation

"Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments... detection and classification of acoustic events may help to detect and describe human activity..." (CLEAR-AED Task Brief)

Difficulties

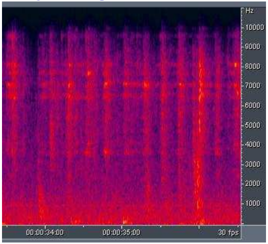
- Negative SNR (speech is "background noise")
- Unknown spectral structure
- Different spectral structure for each event type

Difficulty #1: Negative SNR

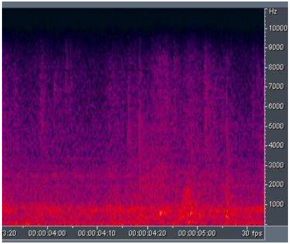


Difficulty #2: Unknown Spectral Structure

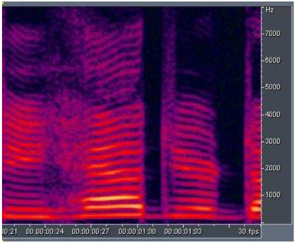
Key Jingle



Footsteps



Speech



Discriminative Feature Selection for AED

Zhuang et al., ICASSP 2008

- **Problem:** what acoustic features are relevant for detecting non-speech acoustic events?
- **Input:** $(x_i \in \Re^D)$ includes many acoustic features invented for speech processing (MFCC, PLP, energy, ZCR)
- **Output:** $(f_i \in \Re^d)$ selects the most useful features:

$$f_i = Wx_i$$

where $W^T = [w_1, \ldots, w_K]$, and w_k is an indicator vector (only one non-zero element)
- **Hidden Markov Modeling:** the label sequence $Y^* = [y_1^*, \ldots, y_N^*]$, $y_i \in \{\text{keyjingle}, \text{footstep}, \ldots\}$ is chosen by a hidden Markov model observing $F = [f_1, \ldots, f_N]$:

$$Y^* = \arg\max p(F|Y)p(Y)$$

Bayes Error Rate

Zhuang et al., ICASSP 2008

Bayes Error Rate

Let w_k be an indicator vector (all zeros except for one element). The Bayes-optimal error rate of a classifier observing feature $w_k^T x$ is

$$P(\text{error}) = \int \int P\left(y \neq \arg\max p(w_k^T x, y)\right) dydx$$

Bayes Error Rate Approximated on a Database

$$\mathcal{F}(w_k) = \frac{1}{N}\sum_{i=1}^{N} \delta\left(y_i \neq \arg\max p(w_k^T x_i, y_i)\right)$$

Feature Selection Algorithms

Hard-Bayes-Error Feature Selection

For $k = 1, \ldots, K$, Choose the indicator vector w_k (w_k is all zeros except for one nonzero element) to minimize

$$\mathcal{F}(w_k) = \frac{1}{N} \sum_{i=1}^{N} \delta \left(y_i \neq \arg\max p(w_k^T x_i, y_i) \right)$$

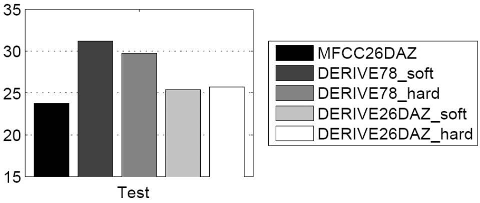
Soft-Bayes-Error Feature Selection

For $k = 1, \ldots, K$, Choose the indicator vector w_k (w_k is all zeros except for one nonzero element) to minimize

$$\mathcal{F}_S(w_k) = \frac{1}{N} \sum_{i=1}^{N} \operatorname{rank} \left(y_i \middle| w_k^T x_i \right)$$

Acoustic Event Detection Results

Zhuang et al., ICASSP 2008



- MFCC26DAZ = 26 Mel-frequency cepstral coefficients + deltas + acceleration
- DERIVE26DAZ = 26 Derived features + deltas + acceleration
- DERIVE78 = 78 Derived features

Discriminative Feature Transform

Work in progress. . .

- **Problem:** what projection of the acoustic spectrogram is relevant for recognizing non-speech acoustic events?
- **Output:** $(f_i \in \Re^d)$ selects the most useful features:

$$f_i = \sum_{k=1}^{K} c_k \sigma(w_k^T x_i)$$

  where $c_k \in \Re^d$ and $w_k \in \Re^D$ are arbitrary real-valued weight vectors, and $\sigma(z) = 1/(1 + e^{-z})$.
- **Hidden Markov Modeling:** the label sequence $Y^* = [y_1^*, \ldots, y_N^*]$, $y_i \in \{\text{keyjingle}, \text{footstep}, \ldots\}$ is chosen by a hidden Markov model observing $F = [f_1, \ldots, f_N]$:

$$Y^* = \arg\max p(F|Y)p(Y)$$

The Baum-Welch Algorithm

Hidden Markov model parameters are trained to maximize the expected log likelihood, with expectation over the unknown state sequence $Q = [q_1, \ldots, q_N]$

$$\mathcal{F} = E_Q \left\{ \log p(F, Q) \right\}$$

$$\mathcal{F} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{q} p(q_i = q | F, Y)(f_i - \mu_q)^T \Sigma_q^{-1} (f_i - \mu_q) - \ldots$$

Baum-Welch Back-Propagation

The neural network can be trained, using standard gradient descent methods, in order to minimize $\mathcal{F}$. For example,

$$f_i = \sum_{k=1}^{K} c_k \sigma(w_k^T x_i)$$

$$\frac{\partial \mathcal{F}}{\partial c_k} = \sum_{i=1}^{N} \sum_{q} p(q_i = q|F) \left(\frac{\partial \mathcal{F}}{\partial f_i} |q_i = q \right) \left(\frac{\partial f_i}{\partial c_k} \right)$$

$$= \sum_{i=1}^{N} \sum_{q} p(q_i = q|F) \Sigma_q^{-1} (\mu_q - f_i) \sigma(w_k^T x_i)$$

The Problem of Spurious Maxima

- It is always possible to train a mixture Gaussian so that $\mathcal{F} = \infty$
  - Solution: Give one of the Gaussians a zero variance ($\Sigma_q = 0$)
  - This is called "over-training"
- In Baum-Welch Back-Propagation, the same result is obtained for $\|c_k\| \rightarrow 0$
- Solution: require $\|c_k\| = 1$, or more generally, $\|\frac{\partial f_i}{\partial x_i}\| = 1$

Methods for Avoiding Spurious Maxima

- Constrained optimization: maximize

$$\mathcal{L} = \mathcal{F} + \sum_k \lambda_k (\|c_k\| - 1)$$

  with Lagrange multipliers λ_k chosen so that $\|c_k\| = 1$
- Symplectic Maximum Likelihood Transform (SMLT, Omar and Hasegawa-Johnson, 2004): replace the neural network with one that computes a *volume preserving* transform:

$$\left|\frac{df}{dx}\right| = 1$$

  where $J_f(x)$ is the Jacobian of the transform

The Reflecting Symplectic Transform

Omar and Hasegawa-Johnson, 2004

Divide x and y arbitrarily into equal-length sub-vectors, $x^T = [x_1^T, x_2^T]$, $y^T = [y_1^T, y_2^T]$. Interpret as follows:

- x_1 is a vector of object positions
- x_2 is a vector of velocities
- $V(x_2)$ is a scalar called the "kinetic energy"
- $T(y_1)$ is a scalar called the "potential energy"
- Then the following transform is volume-preserving:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 - \nabla_{x_2} V \\ x_2 - \nabla_{y_1} T \end{bmatrix} = \begin{bmatrix} x_1 - g_1(x_2) \\ x_2 - g_2(x_1 - g_1(x_2)) \end{bmatrix}$$

- $g_1(x_2)$ and $g_2(y_1)$ must be irrotational. Easiest way to guarantee this: train $V(x_2)$ and $T(y_1)$ directly, using Baum-Welch back-propagation

SMLT+GMM for Phone Classification

Omar and Hasegawa-Johnson, 2004

- Compute phone label y_i given MFCC cepstrum x_i
- Symplectic maximum likelihood transform (SMLT) computes $f_i(x_i)$
- Maximum likelihood linear transform (MLLT) computes $f_i = Wx_i$
- Gaussian mixture model (GMM) computes $p(f_i|y_i)$
- Database: TIMIT

Features	Classifier	Accuracy
MFCC	GMM	73.7%
MLLT	GMM	74.6%
SMLT	GMM	75.6%

Conclusions

- Non-speech acoustic event spectra $\neq$ speech spectra

Conclusions

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection

Conclusions

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection

Conclusions

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection
- Discriminative feature selection can be generalized to discriminative feature transformation

Conclusions

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection
- Discriminative feature selection can be generalized to discriminative feature transformation
- SMLT (a form of discriminative feature transformation) outperforms MFCC and MLLT for phoneme classification in TIMIT

Thank You!

http://www.isle.uiuc.edu/slides/2009/Hasegawa-Johnson09ASA2.pdf