HMM Adaptation for applications in Telecommunication

Karthikesh Raju
Lab. of Comp. & Info. Sc.
karthik@james.hut.fi

2003.03.06
Reference:

Outline

- Recognition Experiments
  - Adaptation to Additive & Convolutional noise
- Application in the Telephone Network
- Conclusions
Outline

- Introduction
- Recognition Experiments
  - Adaptation to Additive & Convolutional noise
- Application in the Telephone Network
- Conclusions
Outline

- Introduction
- Features of the recognizer

- Recognition Experiments
  - Adaptation to Additive & Convolutive noise
- Application in the Telephone Network
- Conclusions
Outline

- Introduction
- Features of the recognizer
- Adaptation of HMMs

- Recognition Experiments
  - Adaptation to Additive & Convolutive noise
- Application in the Telephone Network
- Conclusions
Outline

- Introduction
- Features of the recognizer
- Adaptation of HMMs
  - Estimation of Noise Spectrum
  - Estimation of Frequency Response
  - Adaptation of Cepstral Parameters
- Recognition Experiments
  - Adaptation to Additive & Convolutive noise
- Application in the Telephone Network
- Conclusions
Introduction

- How do we have speech recognition systems in real-life situations?
- **Recognition system at a switch in a telephone network?**
- Recognizer has to cope with two main sources of noise
  - constant background noise (usually additive)
  - Channel noise (usually convolutive)
- Influence of these noise can be described as
  \[ Y(f) = |H(f)|^2 S(f) + N(f) \]
where \( S(f) \) psd of clean speech, \( N(f) \), noise spectrum, \( H(f) \), frequency response of whole transmission system

- This paper assumes a slowly time varying channel

- **Idea:** Adapt HMM parameters with estimates of \( H(f) \) and \( N(f) \)
Features of the recognizer

- Based on representation of speech by cepstral parameters
- Feature vector has
  - 12 MEL frequency cepstral coefficients
  - 12 corresponding $\delta$ cepstral coefficients
- Words are modeled by HMMs with the following features
  - 18 states per word
  - 4 (or 2) Gaussian Mixtures per state
  - left to right model
  - diagonal covariance matrices
Estimation of noise spectrum

- Estimated as a weighted sum of actual and past short-term MEL spectra

\[
\sqrt{\hat{N}(t_i, f)} = \alpha \sqrt{\hat{N}(t_{i-1}, f)} + (1 - \alpha) \sqrt{X(t_i, f)}
\]

- \(\sqrt{\hat{N}(t_i, f)}\) is the estimated magnitude noise spectrum at time \(t_i\)

- Initialization, speech input is preceded only by a background noise segment

- Each sub-band update takes place as long as input spectral component \(\sqrt{X(t_i, f)}\) is below a threshold
- Exceeding the threshold, implies that there is a rise in sub-band energy, which might be due to onset of speech.
- Based on the measurement of

\[ NX(f) = \sqrt{\frac{\hat{N}(t_i, f)}{X(t_i, f)}} \]

- which is noise-to-signal ratio, the authors detect the presence of speech or a non-stationary segment.
- Relative NSRs, indicate if the sub-band has noise or has a speech signal. Speech flag is set if three successive frames indicate the presence of speech.
Presence of speech triggers the HMM adaptation.
Estimation of Frequency Response

\[ |\hat{H}_{act}(f)|^2 = \frac{Y_{long}(f) - \hat{N}(f)}{\hat{S}_{long}(f)} \]

- \( Y_{long}(f) \) long term spectrum assuming a constant \( H(f) \) and a constant \( N(f) \)
- Long term spectra are obtained by summing up various short term spectra in the sub-bands
- Long term spectra of clean speech \( \hat{S}_{long}(f) \), is obtained from the HMM after recognition and alignment.
- Recursively \( |\hat{H}_{act}(f)|^2 \) is estimated as
\[ |\hat{H}_{\text{new}}(f)|^2 = \alpha |\hat{H}_{\text{old}}(f)|^2 + (1 - \alpha) |\hat{H}_{\text{act}}(f)|^2 \]

- Iterative updates results in smoothened version of frequency response
- It also compensates for the estimation errors
- Using the frequency response during the training phase, mismatches between the frequency responses of training and recognition phase can be calculated.
Adaptation of Cepstral Parameters

- Estimates of \( N(f) \), \( H(f) \), modified clean speech spectrum can be calculated.
- This modification requires transformation of the cepstral parameters back to linear spectral domain.
- All cepstral means are adapted against the various noises.
- Adapting only the cepstral parameters - log-add approximation
- The \( \Delta \) cepstral coefficients are adapted by

Robustness ....2003.03.06
\[ \Delta \hat{S}_{lg}(f) \approx \frac{S(f)}{S(f) + \hat{N}(f)} \Delta S_{lg}(f) \]

- \( \Delta S_{lg}(f) \) represents the logarithmic spectral parameters when transforming back the \( \Delta \) cepstral coefficients.
Recognition Experiments

- Speaker independent recognition of digit sequences and isolated digits
- TIDIGITS data base are used for training
- Original data recorded at a high SNR
- Each digit is modeled by a single HMM consisting of a mixture of four Gaussian components per state
- Recognition is done with adding noise to the TIDIGITS and filtering them
- A Bellcore database consists of isolated digits recorded via telephone lines
Additive Noise Adaptation

- Artificial addition of car noise to the TIDIGITS at different SNR
- Results without noise are at 30dB
- Without adaptation error rate increases, while Delta adaptation further reduces error rates
- Adaptation of Delta Coefficients and variances - worthwhile only at low SNR (variance adaptation is an computationally expensive process)
Fig. 6. Word error rates applying the Log-add approximation.
Additive Noise Adaptation -2

- PMC (Parallel Model Combination) and Spectral Subtraction (SS)
- SS can be used as a preprocessing method that adds robustness to the recognizer
- SS alone produces significant improvement, but HMM adaptation improves the error rates further
- Leads to the hypothesis: SS might introduce certain distortions.
- **Segments with low energy and spectrally similar to noise might get attenuated which might have negative effects on the recognition process**
Other Noise sources are helicopter noise and speech-like-noise
Fig. 7. Word error rates comparing the Log-add approximation against spectral subtraction.
Additive and Convolutive Noise Adaptation

- Test data is filtered with a frequency characteristic simulating a telephone channel ($f < 300$ Hz & $f > 3400$ Hz attenuated by 40 dB)
- Amplification of 3dB/octave for the range 300 to 1000 Hz
- **without adaptation:** 4.23 %
- **LA with filter estimation:** 0.71 %
connected TIDIGITS recognition

vocabulary: 1-9, zero, oh
car noise artificially added
to the filtered data
8700 test utterances with
a total of 28583 digits

Fig. 9. Word error rates when recognizing the filtered data with
car noise added.

Log-add approx.
with noise & filter
estimates

without adaptation
Telephone Network Adaptation

- Field test using a real network
- Callers utter *yes, no, results of some additions*
- PC connected to the ISDN line
- 2-gender dependent HMMs, trained from a data base of German and Swedish people speaking English
  - **without adaptation**: 7.98 %
  - **with adaptation**: 3.47 %
Conclusions

- Adaptation of HMMs to
  - stationary background noise
  - Frequency mismatch between training sequences and test data
- Processing is based on PMC approaches where noise spectrum as well as frequency response are estimated
- Considerable improvements can be gained by modeling garbages (breathing before and after the speech etc)
- Results show good gains with adaptation
Results are based on adaptation of the cepstral means, adaptation of the distributions is more complex (computationally), but should invariably lead to better results.

Goal: integrate the recognition system to a telephone network.