Effects of Frequency-Based Inter-frame Dependencies on Automatic Speech Recognition

Ludovic Trottier Brahim Chaib-draa Philippe Giguère

ludovic.trottier.1@ulaval.ca
{chaib,philippe.giguere}@ift.ulaval.ca

Laval University

May 8, 2014

Outline

Outline

Automatic Speech Recognition

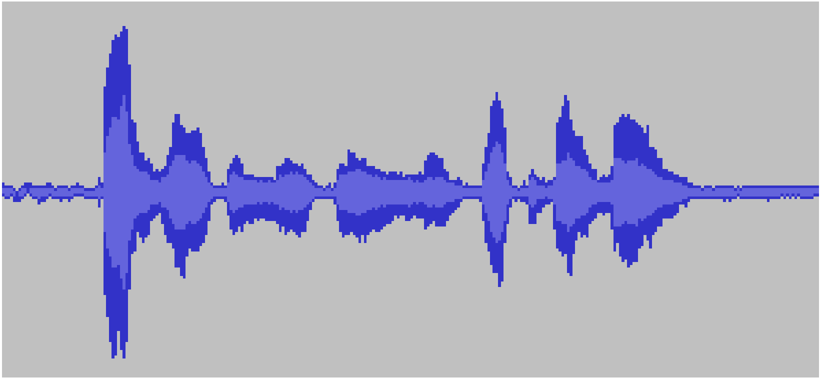
Automatic Speech Recognition

Automatic speech recognition tries to solve problems like :

Automatic Speech Recognition

Automatic speech recognition tries to solve problems like :

what is the person saying?

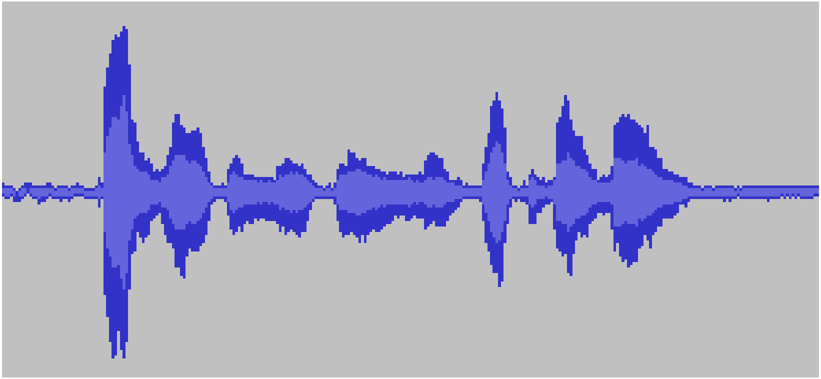


Answer :

Automatic Speech Recognition

Automatic speech recognition tries to solve problems like :

what is the person saying?

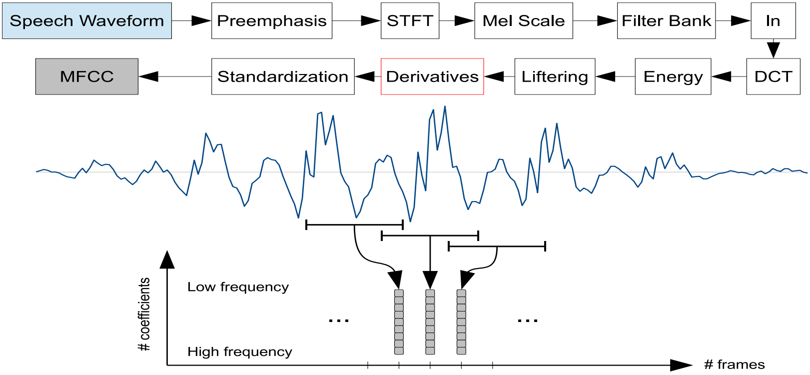


Answer : All work and no play makes Jack a dull boy

Features

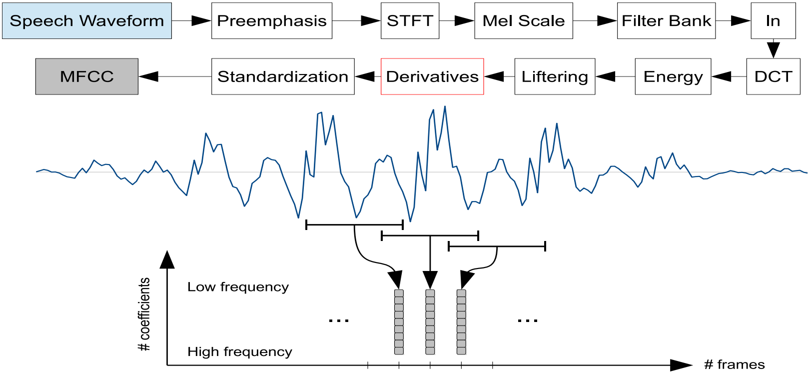
Features

MFCC : Mel-frequency cepstral coefficients



Features

MFCC : Mel-frequency cepstral coefficients



The differentiation of a noisy signal amplifies the noise. Can something else be used?

Contributions

Contributions

1. **Features**

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations
   Signal processing theories show that the rate at which information changes in signals is proportional to frequency.

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations
   Signal processing theories show that the rate at which information changes in signals is proportional to frequency.

2. **Model**

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations
   Signal processing theories show that the rate at which information changes in signals is proportional to frequency.

2. **Model**
   A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations
   Signal processing theories show that the rate at which information changes in signals is proportional to frequency.

2. **Model**
   A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.
   Motivations

Contributions

1. **Features**
   In place of derivatives, we used coefficients concatenation based on the time and the frequency.
   Motivations
   Signal processing theories show that the rate at which information changes in signals is proportional to frequency.

2. **Model**
   A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.
   Motivations
   High-dimensional features.

Triangular Window (Contribution 1)

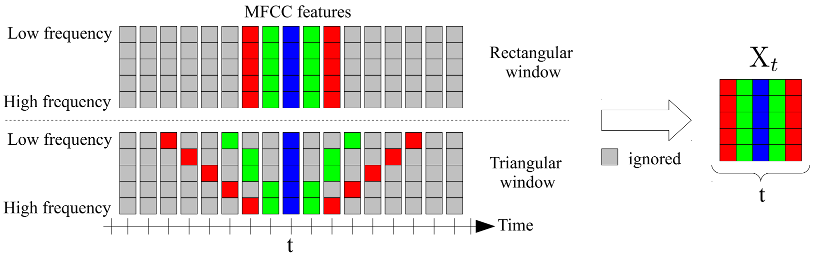
Triangular Window (Contribution 1)

Original MFCC

Concatenation of derivatives

Our Features

Concatenation of coefficients on each side of a frame according to the shape of the window.



Motivations for Triangular Window

Motivations for Triangular Window

Variation of the intensity of different frequency components.

Long and continuous lines implies slow variation.

Short lines implies high variation.

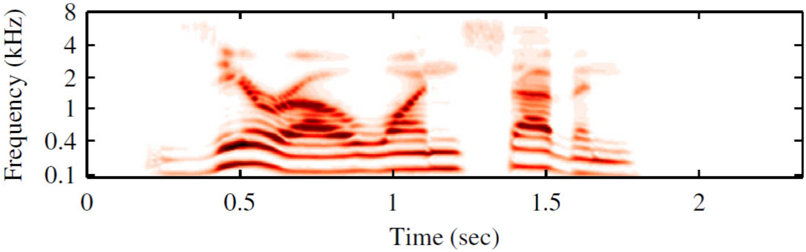
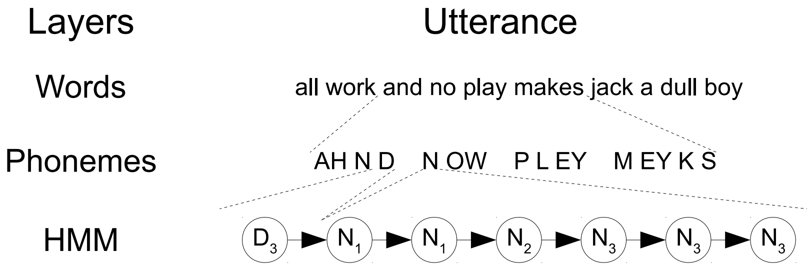


Figure taken from [Heckmann et al., 2011]

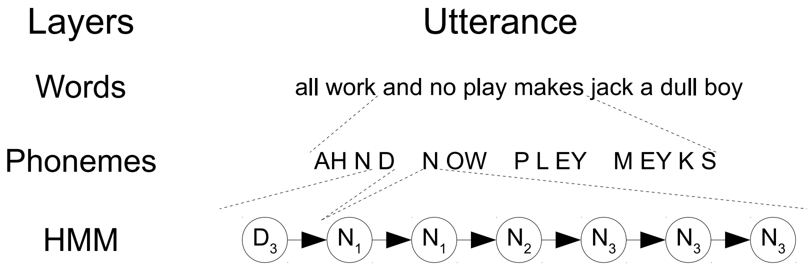
Outline

Phoneme Based Hidden Markov Model

Phoneme Based Hidden Markov Model



Phoneme Based Hidden Markov Model



Original Model

HMM with a mixture of Gaussian densities. (GMM-HMM)

Our Model

HMM with a mixture of Matrix Normal densities. (MNMM-HMM)

Learning MNMM-HMM (Contribution 2)

Learning MNMM-HMM (Contribution 2)

Matrix Normal Distribution

Let X and M be $n \times p$ dimensional matrices, U be $n \times n$ and V be $p \times p$. If $X \sim \mathcal{MN}(M, U, V)$, then:

Learning MNMM-HMM (Contribution 2)

Matrix Normal Distribution

Let X and M be $n \times p$ dimensional matrices, U be $n \times n$ and V be $p \times p$. If $X \sim \mathcal{MN}(M, U, V)$, then:

$$p(X|M, U, V) = \frac{\exp\left(-\frac{1}{2}\,\mathrm{Tr}\left[V^{-1}(X - M)^{\top}U^{-1}(X - M)\right]\right)}{(2\pi)^{\frac{np}{2}}|V|^{\frac{n}{2}}|U|^{\frac{p}{2}}}\ , \quad (1)$$

where M is the mean, U is the among-row variance and V the among-column variance.

Learning MNMM-HMM (Contribution 2)

Matrix Normal Distribution

Let X and M be $n \times p$ dimensional matrices, U be $n \times n$ and V be $p \times p$. If $X \sim \mathcal{MN}(M, U, V)$, then:

$$p(X|M, U, V) = \frac{\exp\left(-\frac{1}{2}\operatorname{Tr}\left[V^{-1}(X-M)^\top U^{-1}(X-M)\right]\right)}{(2\pi)^{\frac{np}{2}}|V|^{\frac{n}{2}}|U|^{\frac{p}{2}}}, \quad (1)$$

where M is the mean, U is the among-row variance and V the among-column variance.

Learning

Using the posterior probability computed by the well-known Forward-Backward recursion, we can update the parameters M, U and V.

Outline

Aurora 2

Aurora 2

Aurora 2 task is :

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses
- Noisy : SNR between -5 and 20 dB

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses
- Noisy : SNR between -5 and 20 dB
- Train : 16,880 utterances

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses
- Noisy : SNR between -5 and 20 dB
- Train : 16,880 utterances
- Test A : 28,028 utterances

Aurora 2

Aurora 2 task is :

- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses
- Noisy : SNR between -5 and 20 dB
- Train : 16,880 utterances
- Test A : 28,028 utterances
- Test B : 28,028 utterances

Aurora 2

Aurora 2 task is :

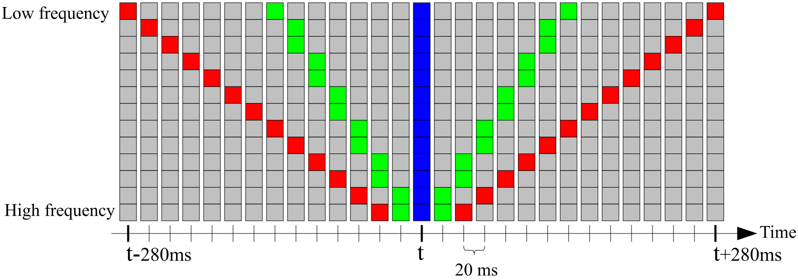
- 11 spoken digits : *zero* to *nine* with *oh*
- Connected speech: any order, up to 7, possible pauses
- Noisy : SNR between -5 and 20 dB
- Train : 16,880 utterances
- Test A : 28,028 utterances
- Test B : 28,028 utterances
- Test C : 14,014 utterances

Triangular Window

Triangular Window

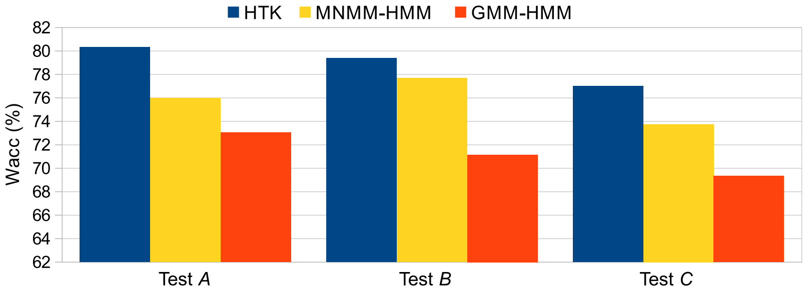
We tested multiple configurations for the triangular window.

This is the shape of the best triangular window we found.



Triangular Window (cont.)

Triangular Window (cont.)



Our features, our model

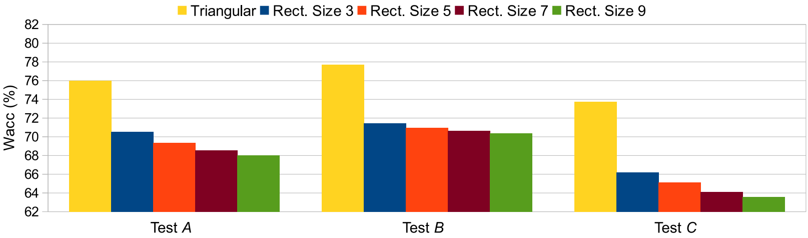
Original features, original model

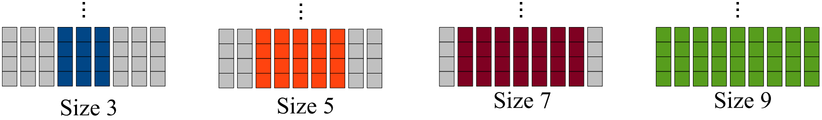
Reference only

Our implementation shows that its helps simpler speech recognizer.

Rectangular VS Triangular

Rectangular VS Triangular





Taking into account the frequency of the coefficients, the triangular window outperformed mere concatenation.

Adding more information degrades the performances of rectangular.

Conclusion

Contributions

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Results

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Results

The concatenation of adjacent features might be a better idea than the concatenation of derivatives.

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Results

The concatenation of adjacent features might be a better idea than the concatenation of derivatives.

Future Work

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Results

The concatenation of adjacent features might be a better idea than the concatenation of derivatives.

Future Work

1. Feature representation model to learn the shape of the window.

Conclusion

Contributions

1. In place of derivatives, we used coefficients concatenation based on the time and the frequency.
2. A Hidden Markov Model with a Matrix Normal Mixture Model as the emission density was designed.

Results

The concatenation of adjacent features might be a better idea than the concatenation of derivatives.

Future Work

1. Feature representation model to learn the shape of the window.
2. Triangular window directly on the waveform.