Speech Recognition or Speech-to-Text conversion: The first block of a virtual character system.

Panos Georgiou
Research Assistant Professor (Electrical Engineering)
Signal and Image Processing Institute
http://sail.usc.edu

State of the Art in Speech Recognition

- LVCSR = Large Vocubulary Speech Recognition
- ASR = Automatic Speech Recognition
- WER = Word Error Rate (can be above 100%)

- Current state of the art error rates range dramatically by task:

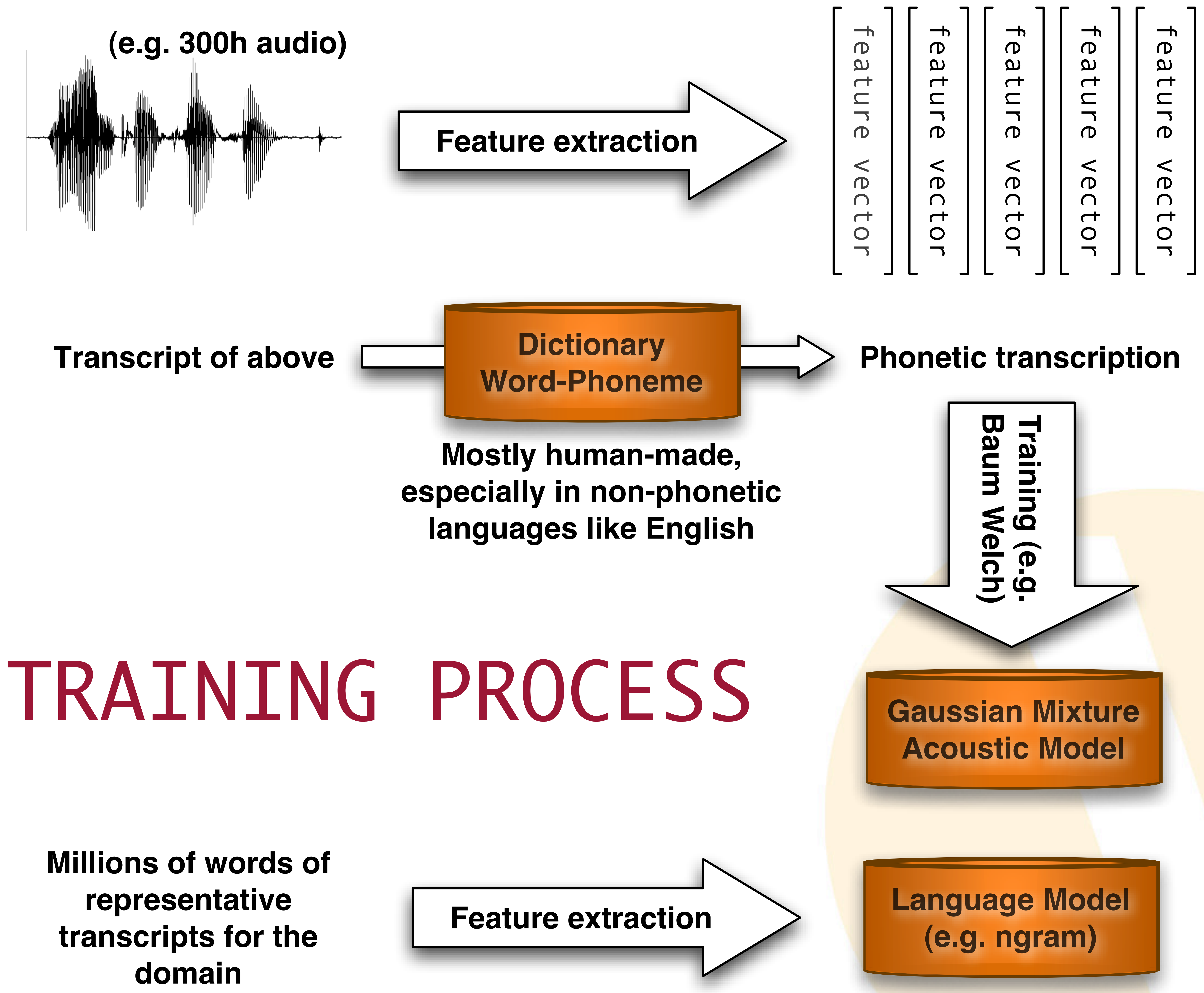
(not all are real time systems)

Digits	11	0.5
Read speech (WSJ)	5K	3
Read speech (WSJ)	20K	3
Broadcast news	64K	10
Conversational telephone	64K	20

Virtual character?

Data starved
1K seen, but 15K models
models

Speech Recognition: Training Process



Speech Recognition: Recognition (Decoding) Process

- **Decoding:**
  - Word sequence = the word sequence that is maximum given the observations

$$\hat{W} = \arg\max_{W \in D} P(W|O)$$

  - It is mathematically the same as (Bayes rule)

$$\hat{W} = \arg\max_{W \in D} \frac{P(O|W)P(W)}{P(O)}$$

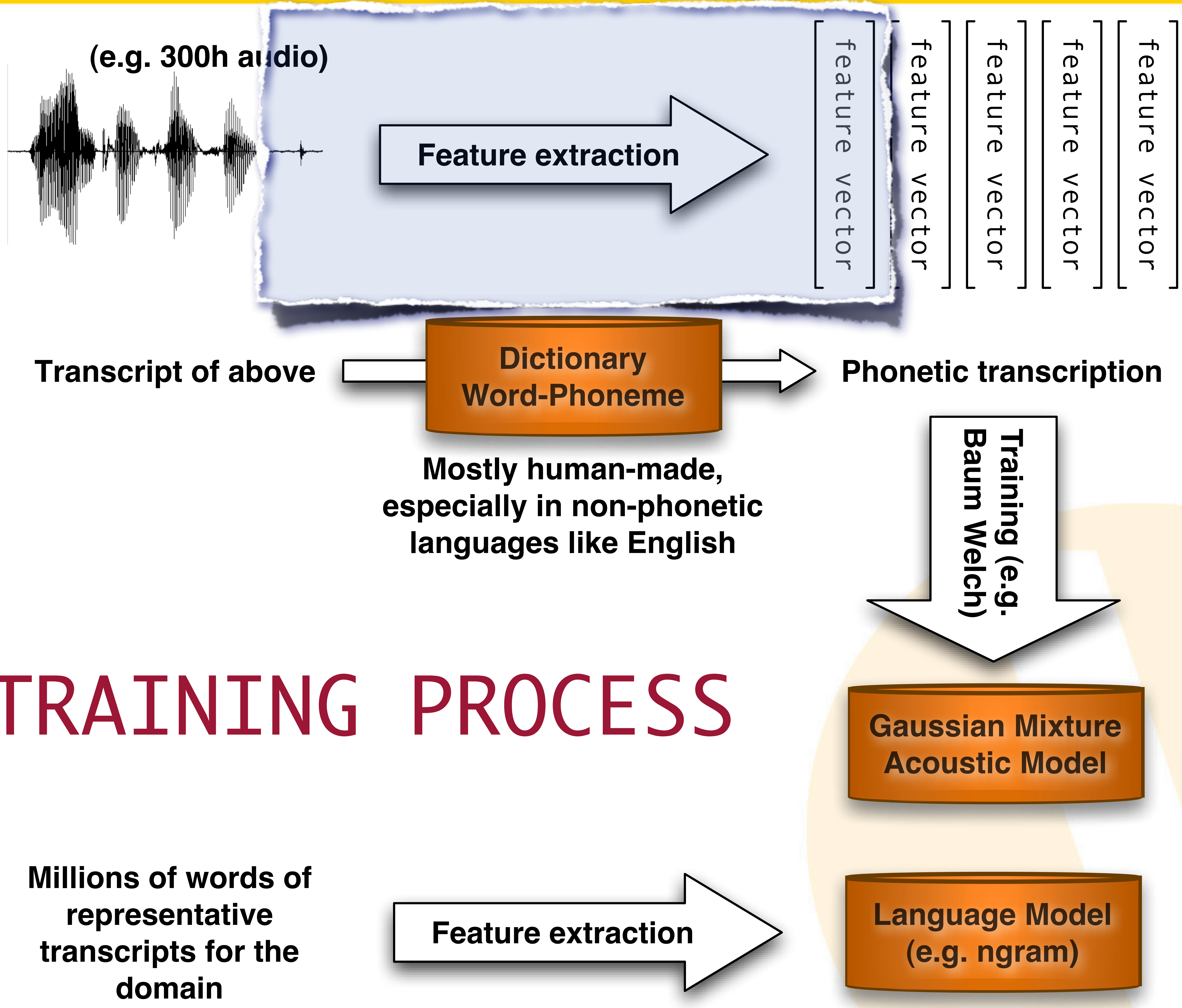
  - And we can drop the common denominator

$$\hat{W} = \arg\max_{W \in D} \underbrace{P(O|W)}_{\text{Acoustic Model}}\underbrace{P(W)}_{\text{Language Model}}$$

  - Real life:

$$\hat{W} = \arg\max_{W \in D} P(O|W)P(W)^N$$

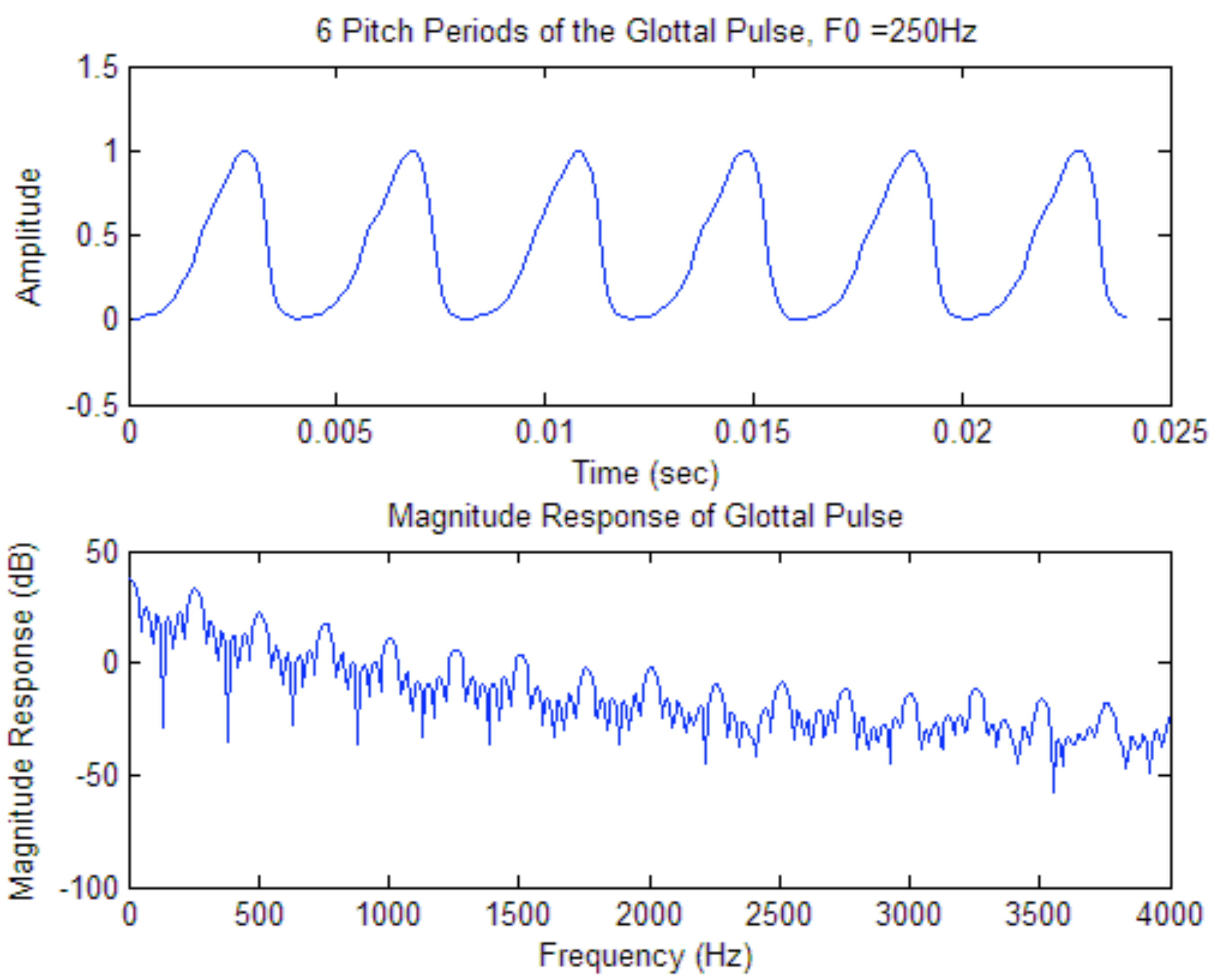
Speech Recognition: Training Process

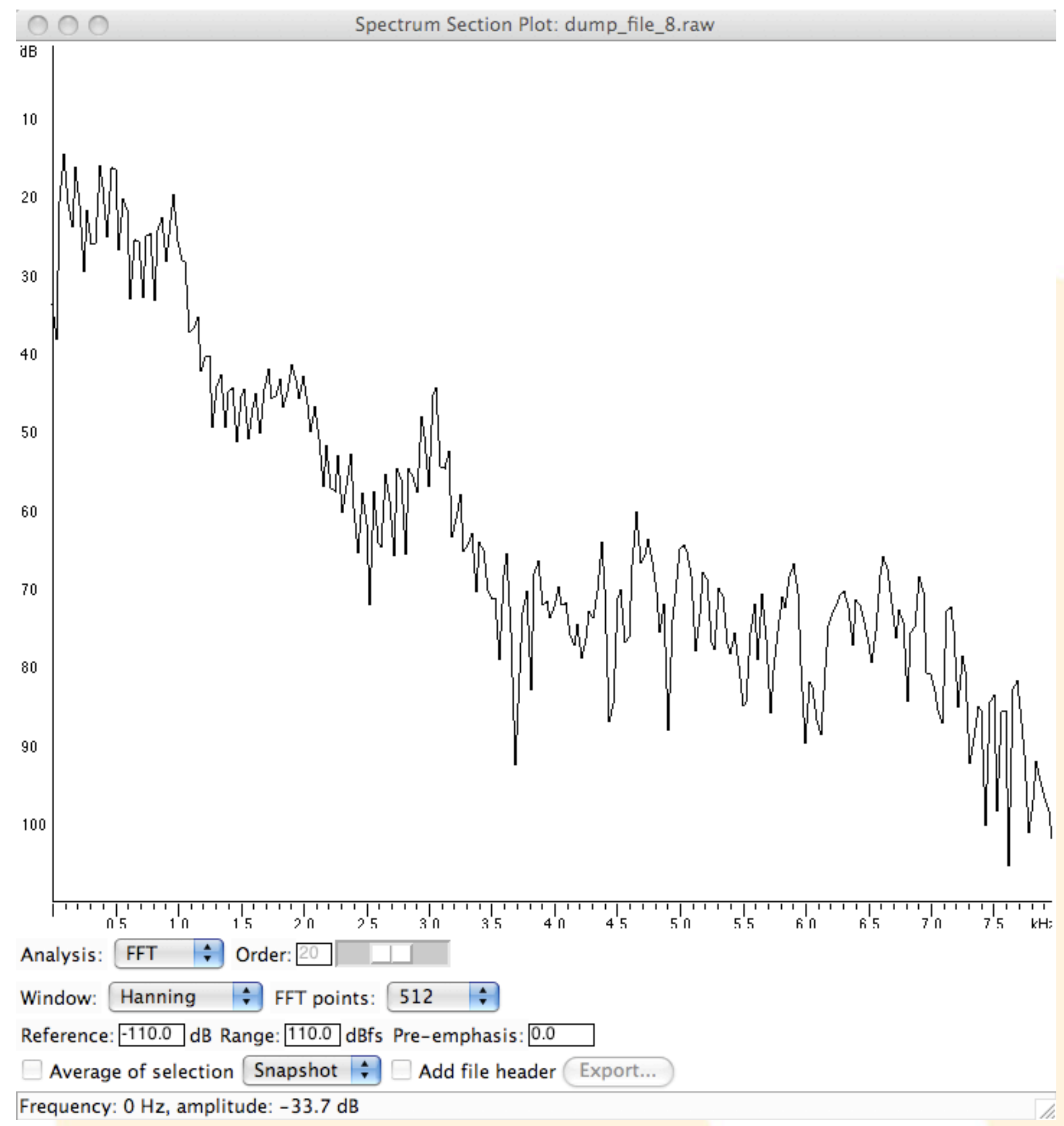


TRAINING PROCESS

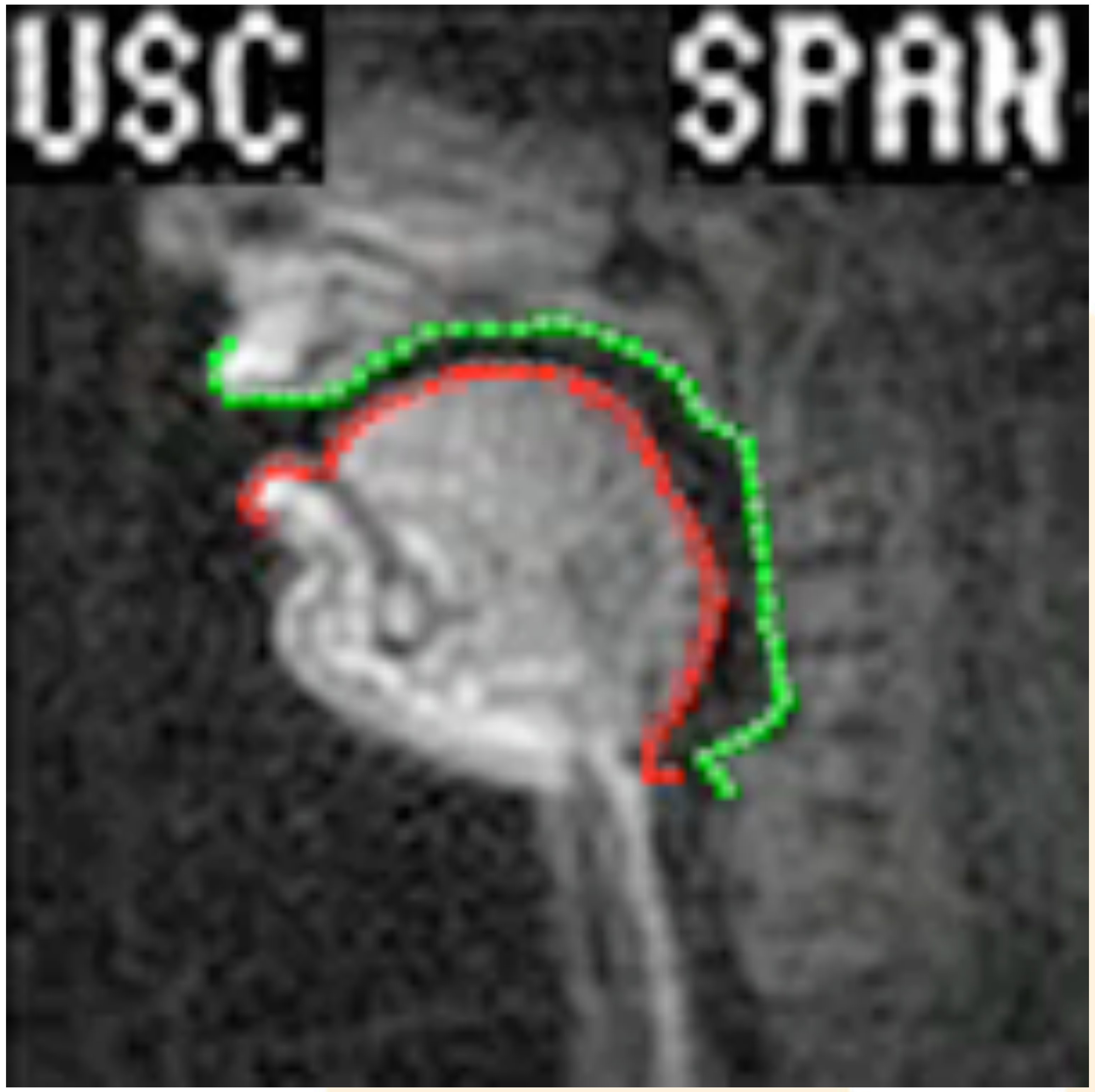
Feature Characterization

- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a 'tube'





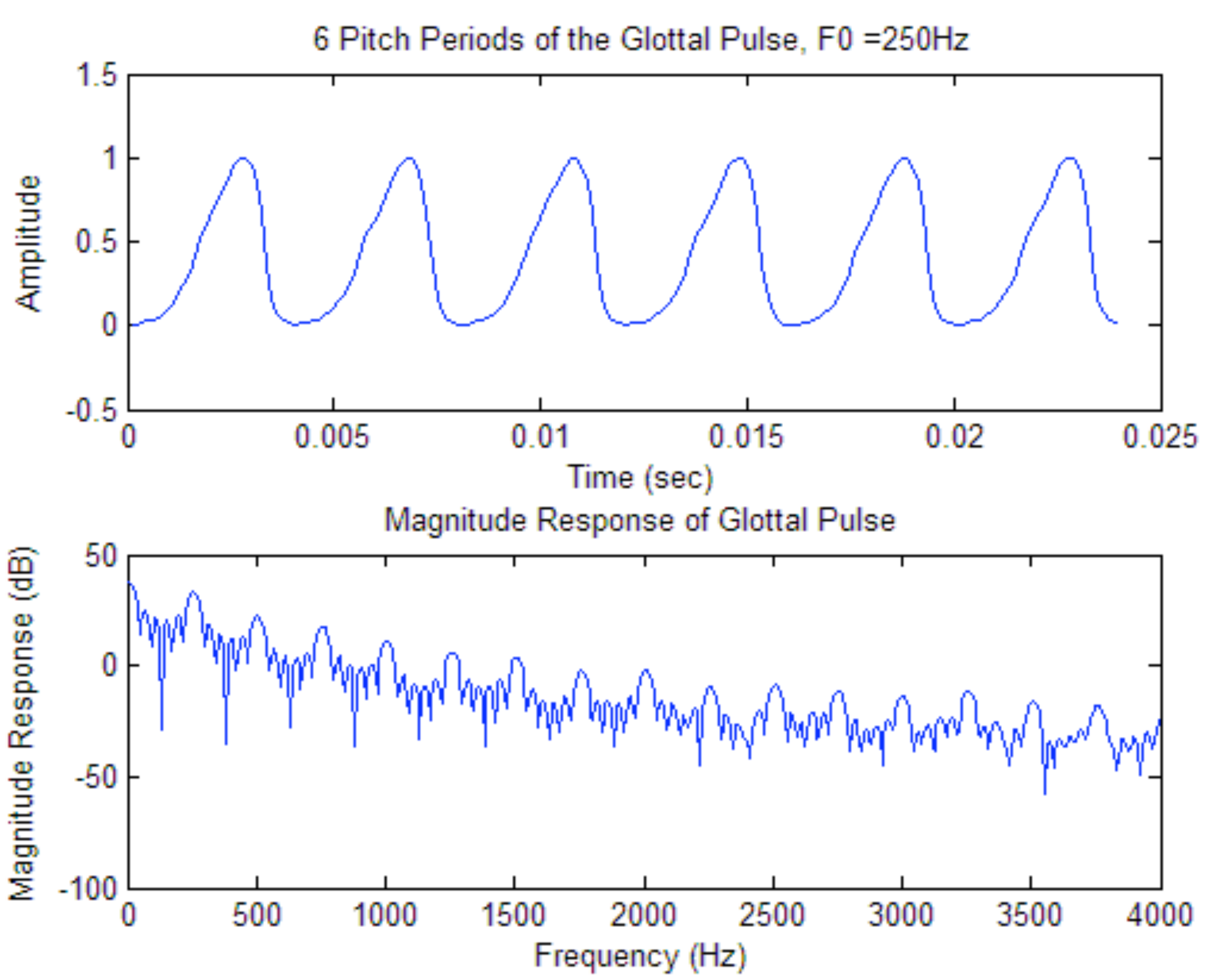
- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a 'tube'



Feature Characterization

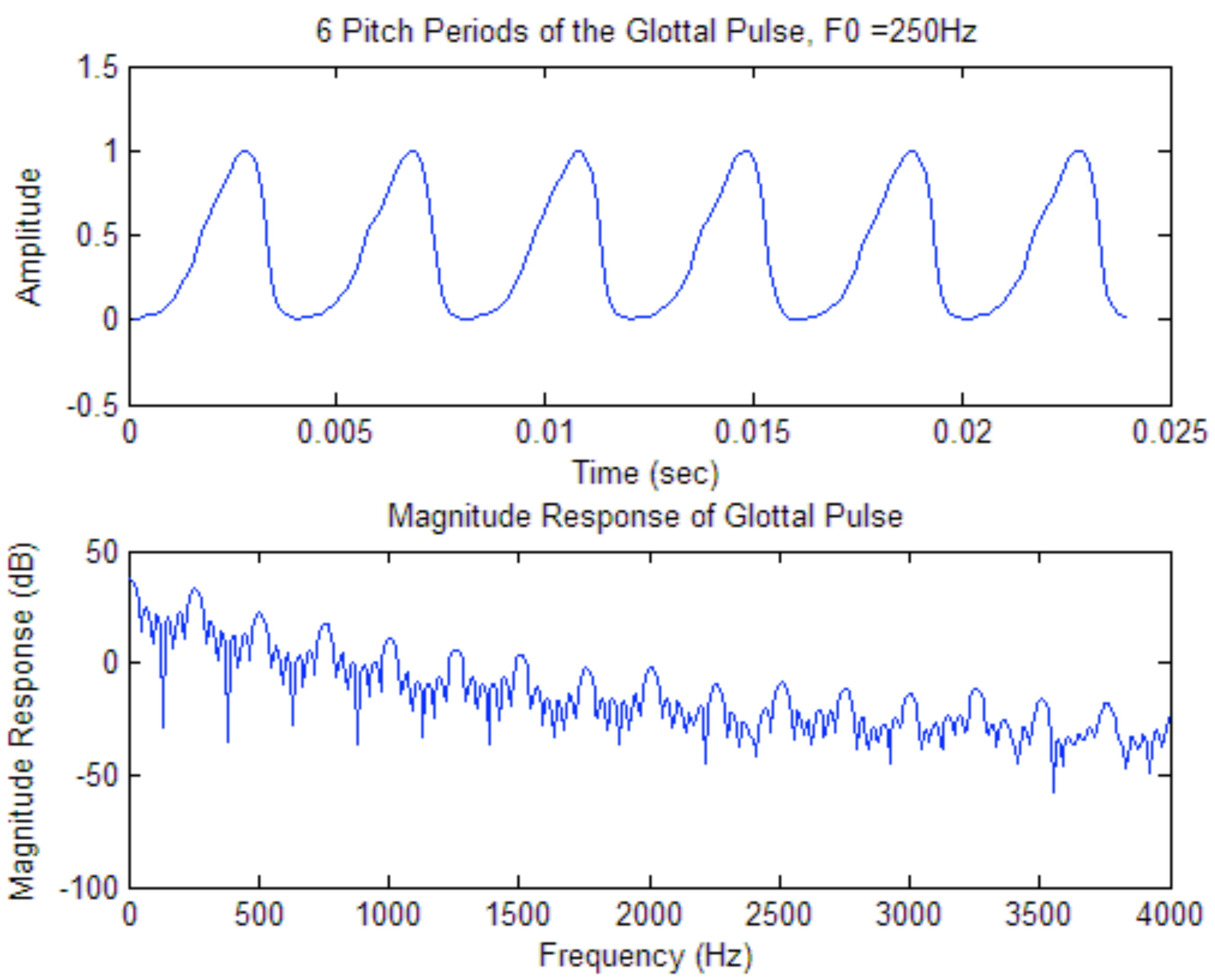
- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a 'tube'

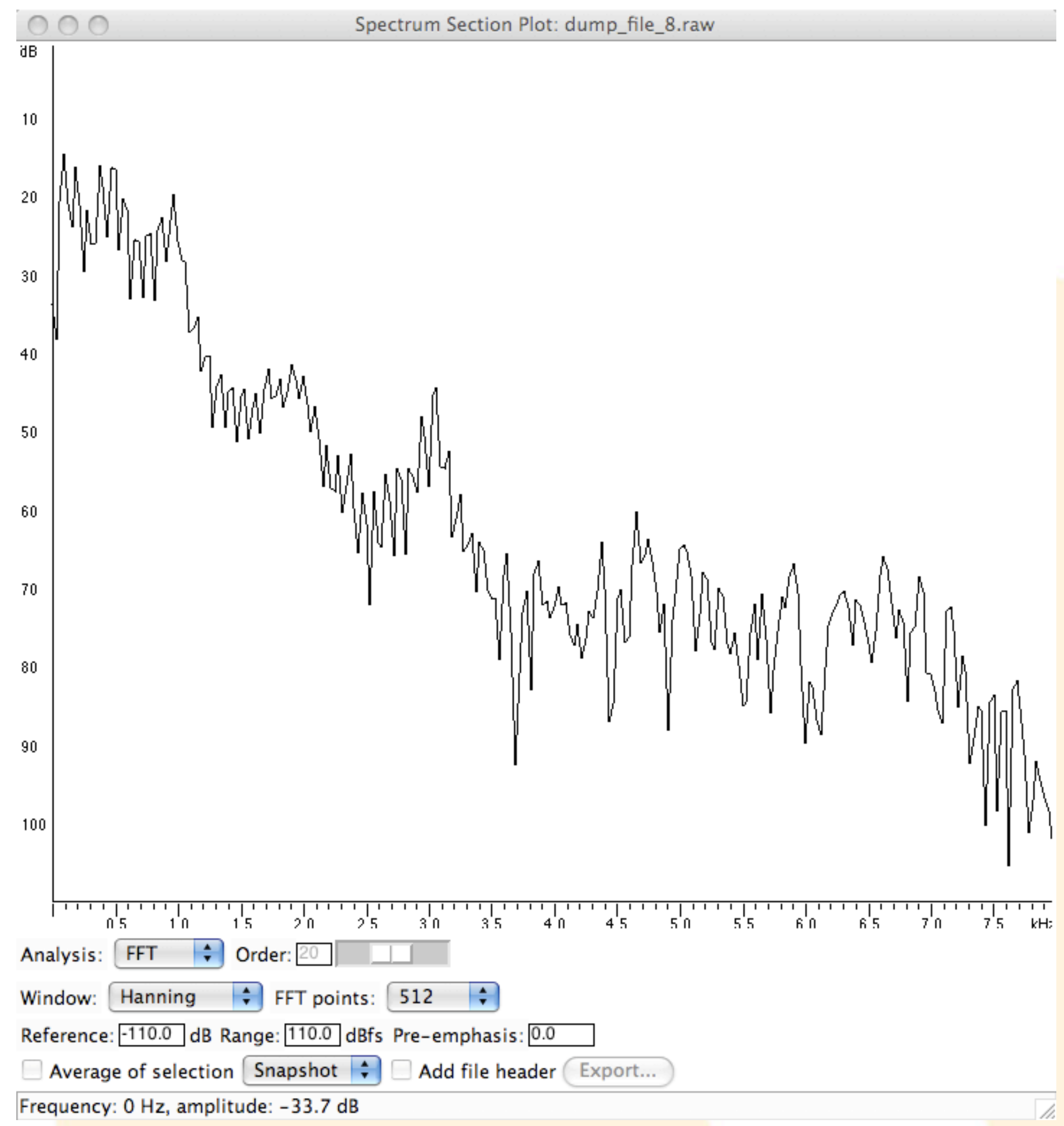
- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a ‘tube’



Feature Characterization

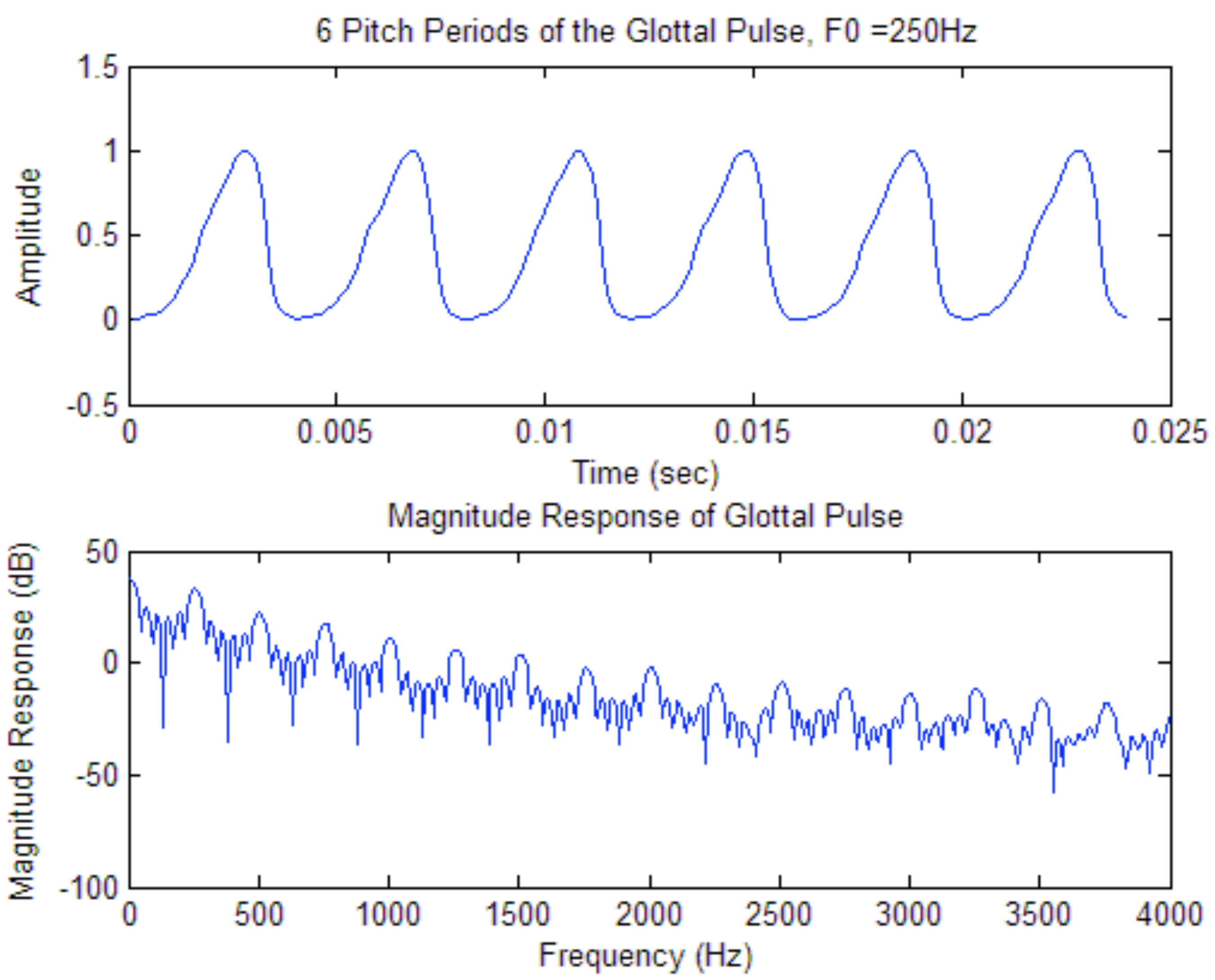
- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a ‘tube’

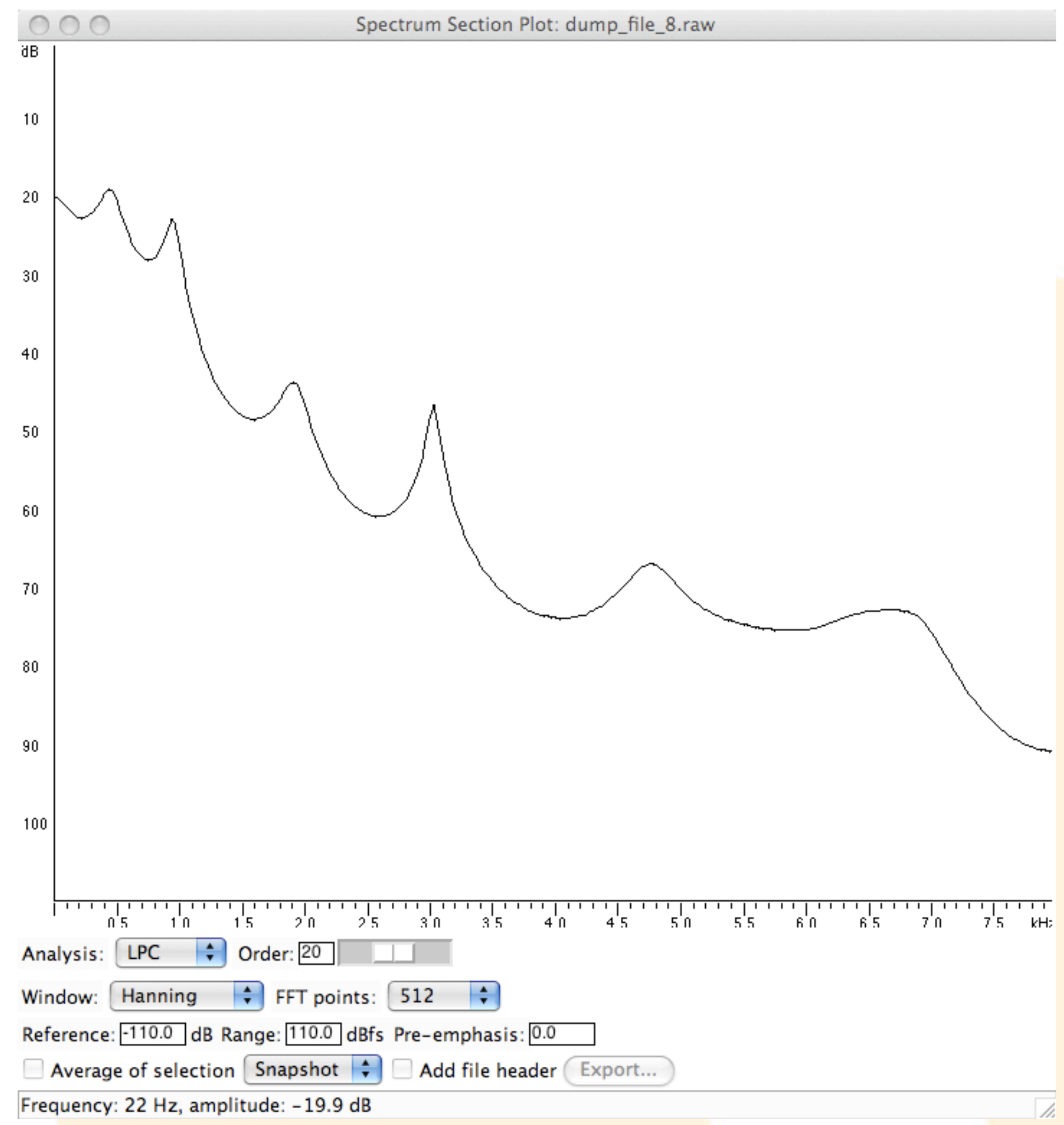




Feature Characterization

- **Acoustic representation:**
  - In short take advantage of spectral characteristics
  - Think of voiced sounds like harmonics of the vocal chord vibrations, that due to shape of the vocal tract create resonances. Different sounds, different resonances
  - Early work approximates the vocal tract with a 'tube'

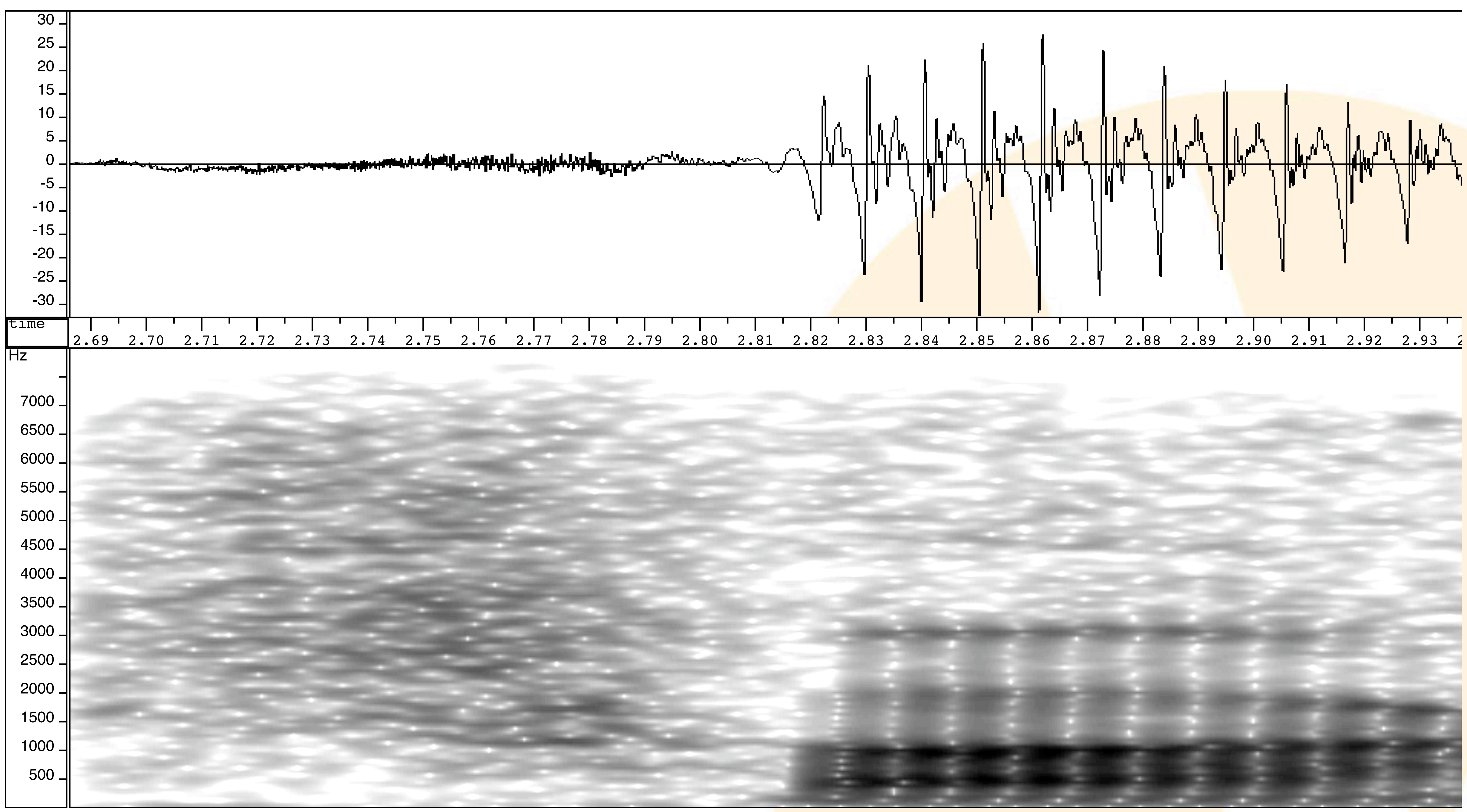




Features

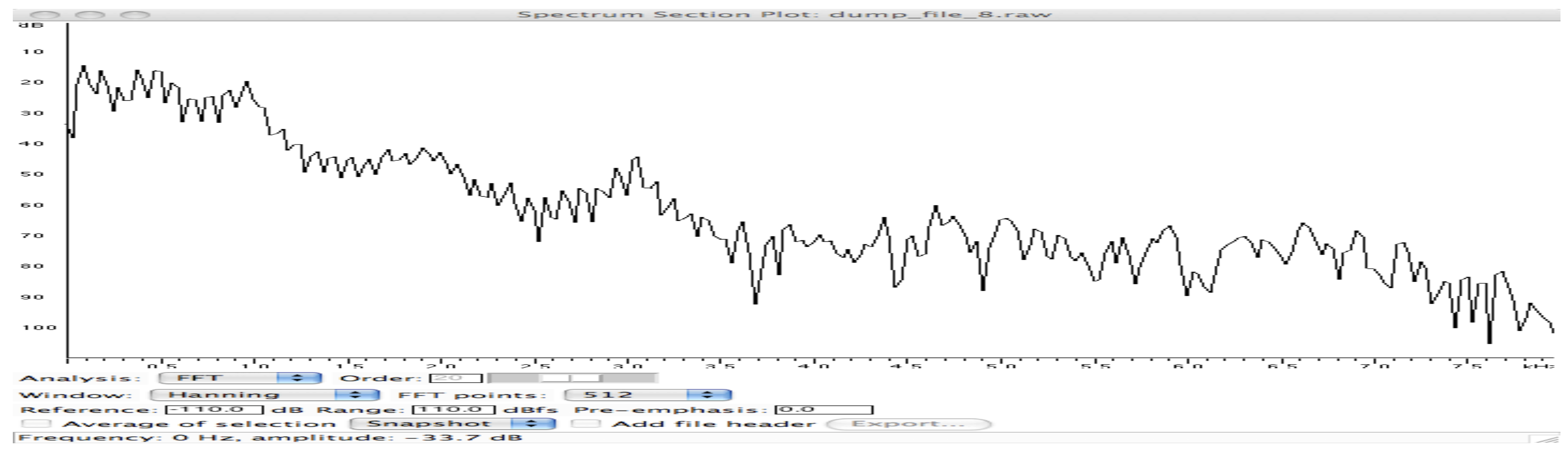
- **Acoustic representation:**
  - Speech signal complex, with fricatives, voiced, unvoiced, plosives etc....
  - Spectrum good for visualizing voiced sounds
  - LPC (last slide) one option.

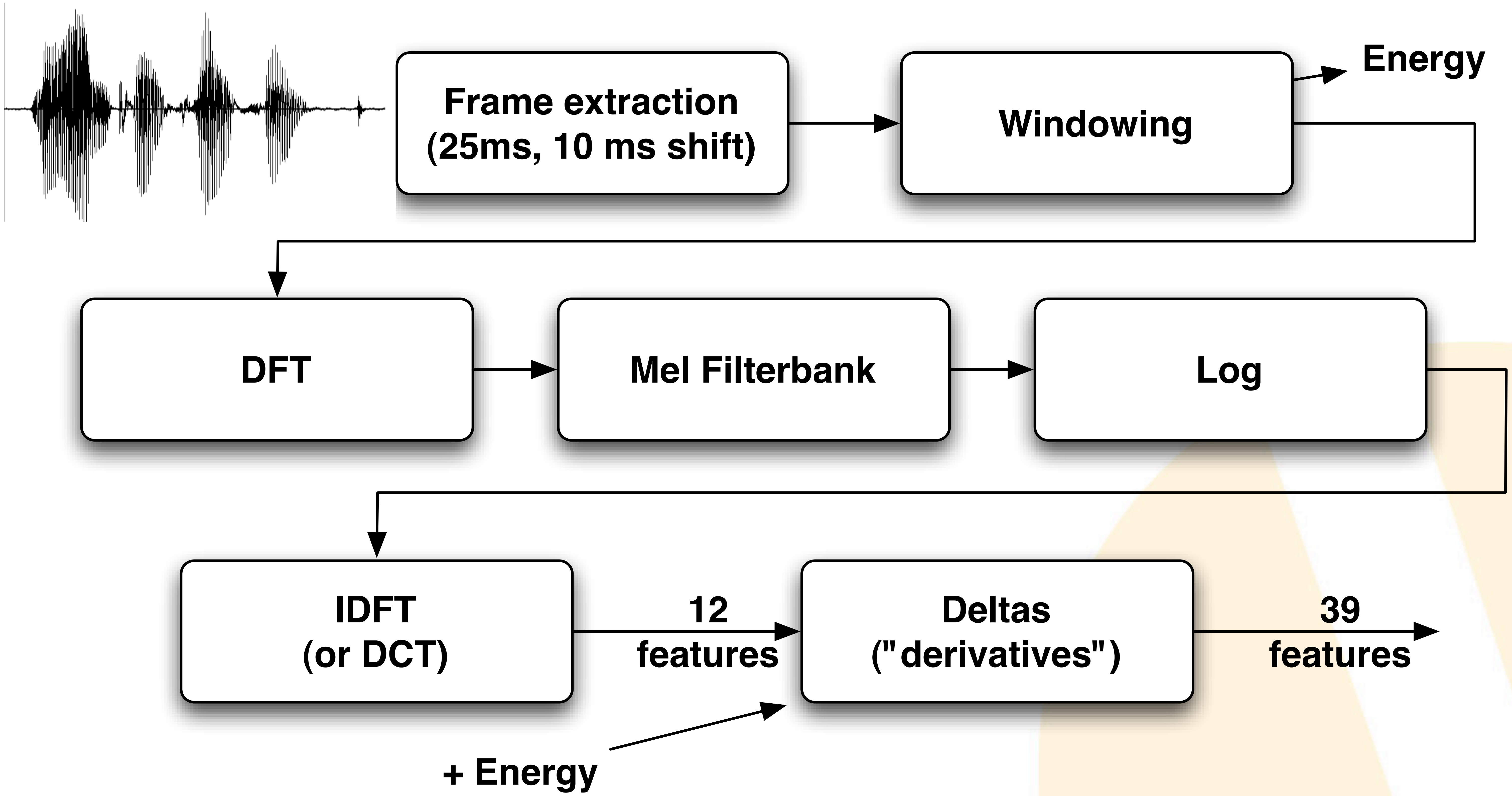
File: /Users/georgiou/cmusphinx/OtoSenseP/dump_file_8.raw   Page: 1 of 2   Printed: Fri Sep 19 02:30:54



Features: Mel Frequency Cepstral Coefficients

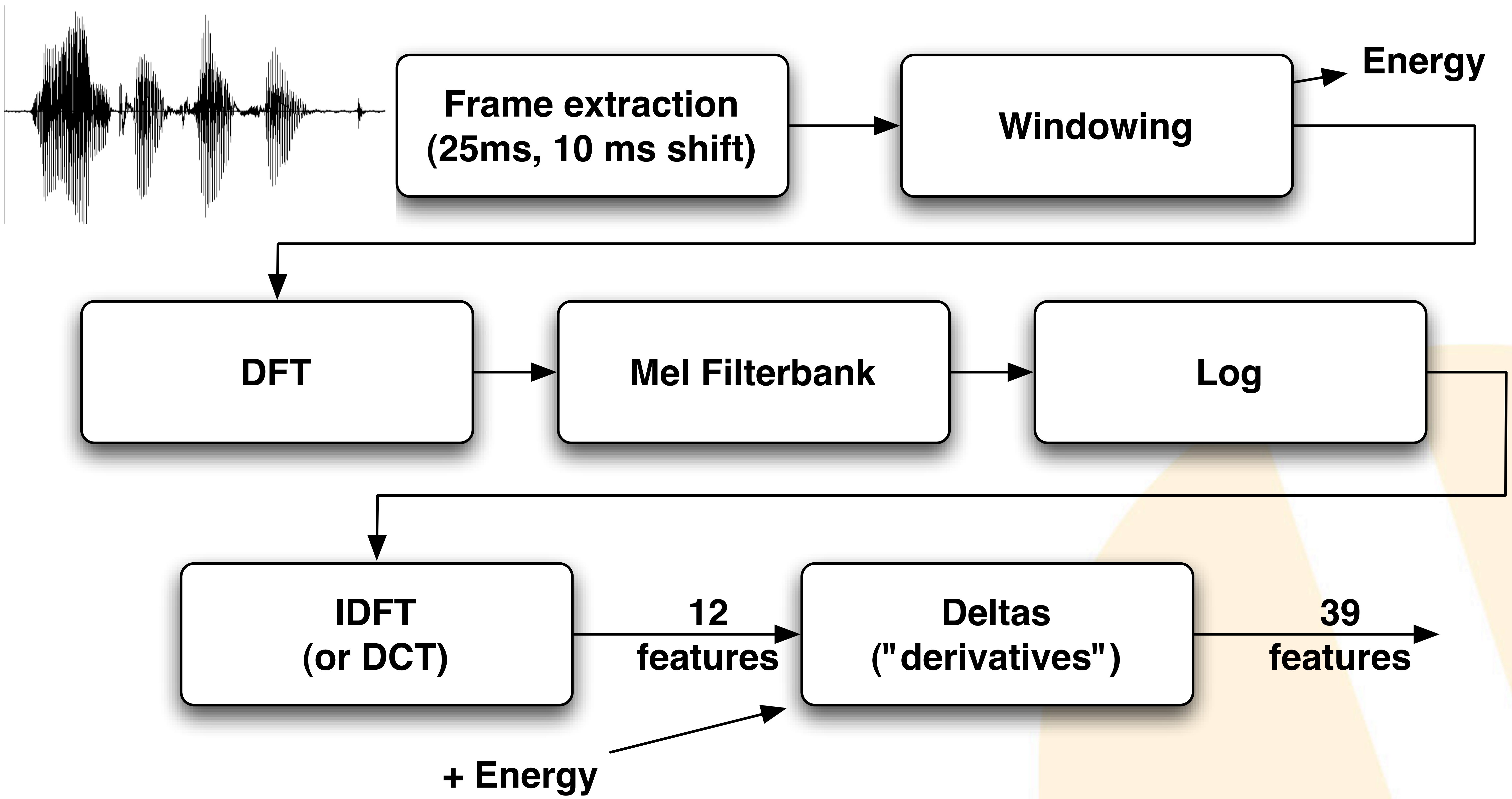
- **More commonly than LPC:**
- **MFCC = Mel Frequency Cepstral Coefficients**



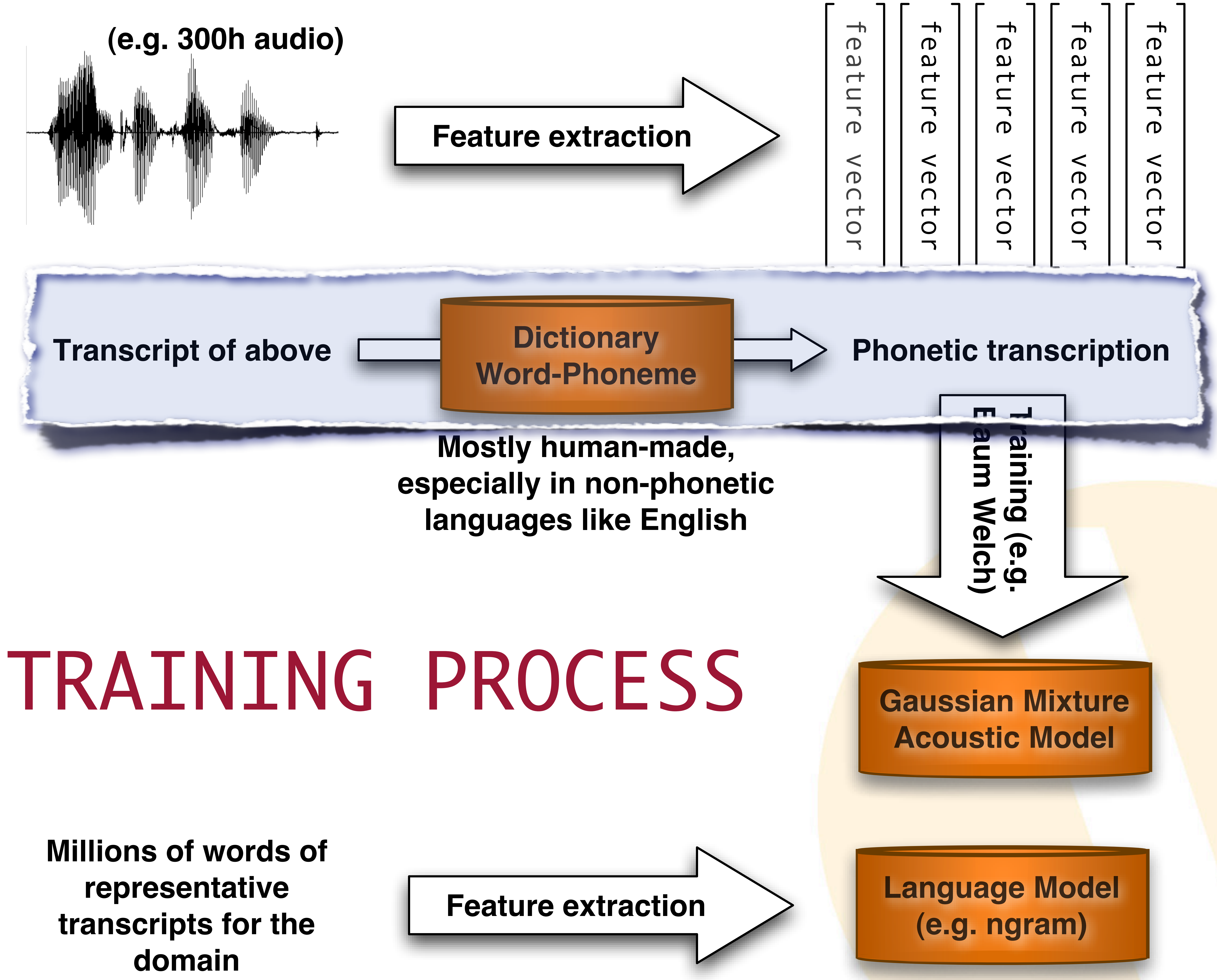


Features: Mel Frequency Cepstral Coefficients

- **More commonly than LPC:**
- **MFCC = Mel Frequency Cepstral Coefficients**

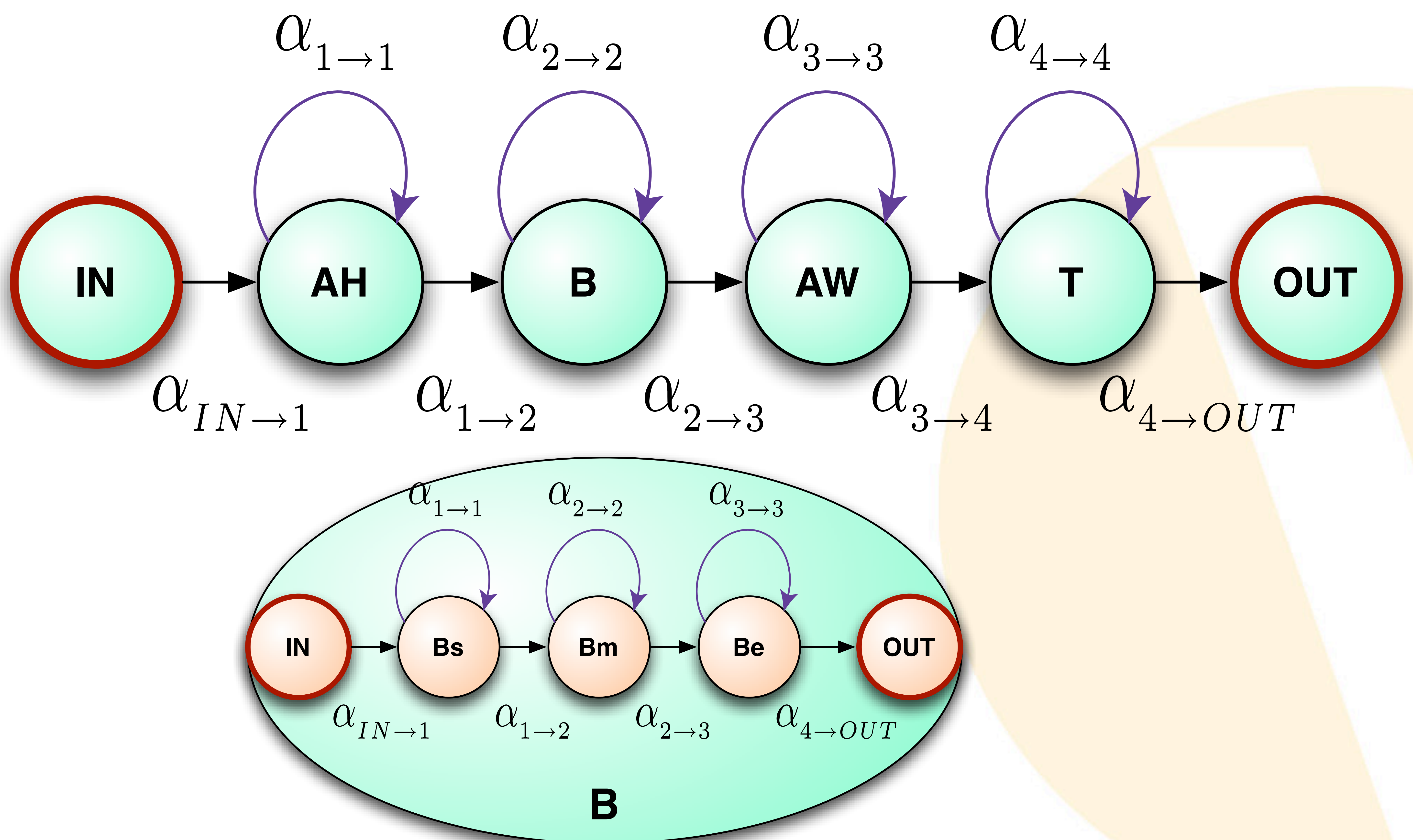


Speech Recognition: Training Process



Components: Lexicon or Dictionary

- **In simple representation:**
  - ABOUT AH B AW T
  - ABSORPTION AH B S AO R P SH AH N
  - ABSORPTION(2) AH B Z AO R P SH AH N
- **But in reality each of these are an Hidden Markov Model state:**



Components: Lexicon or Dictionary

- **In reality it is more complicated**
- **We use triphone models**
  - ABOUT $\_AH_{B}$ $_{AH}\mathbf{B}_{AW}$ $_{B}AW_{T}$ $_{AW}T\_$
  - ABSORPTION $\_AH_{B}$ $_{AH}\mathbf{B}_{S}$ $_{B}S_{AO}$ $_{S}AO_{R}$ $_{AO}R_{P}$ $_{R}P_{SH}$ $_{P}SH_{AH}$ $_{SH}AH_{N}$ $_{AH}N\_$

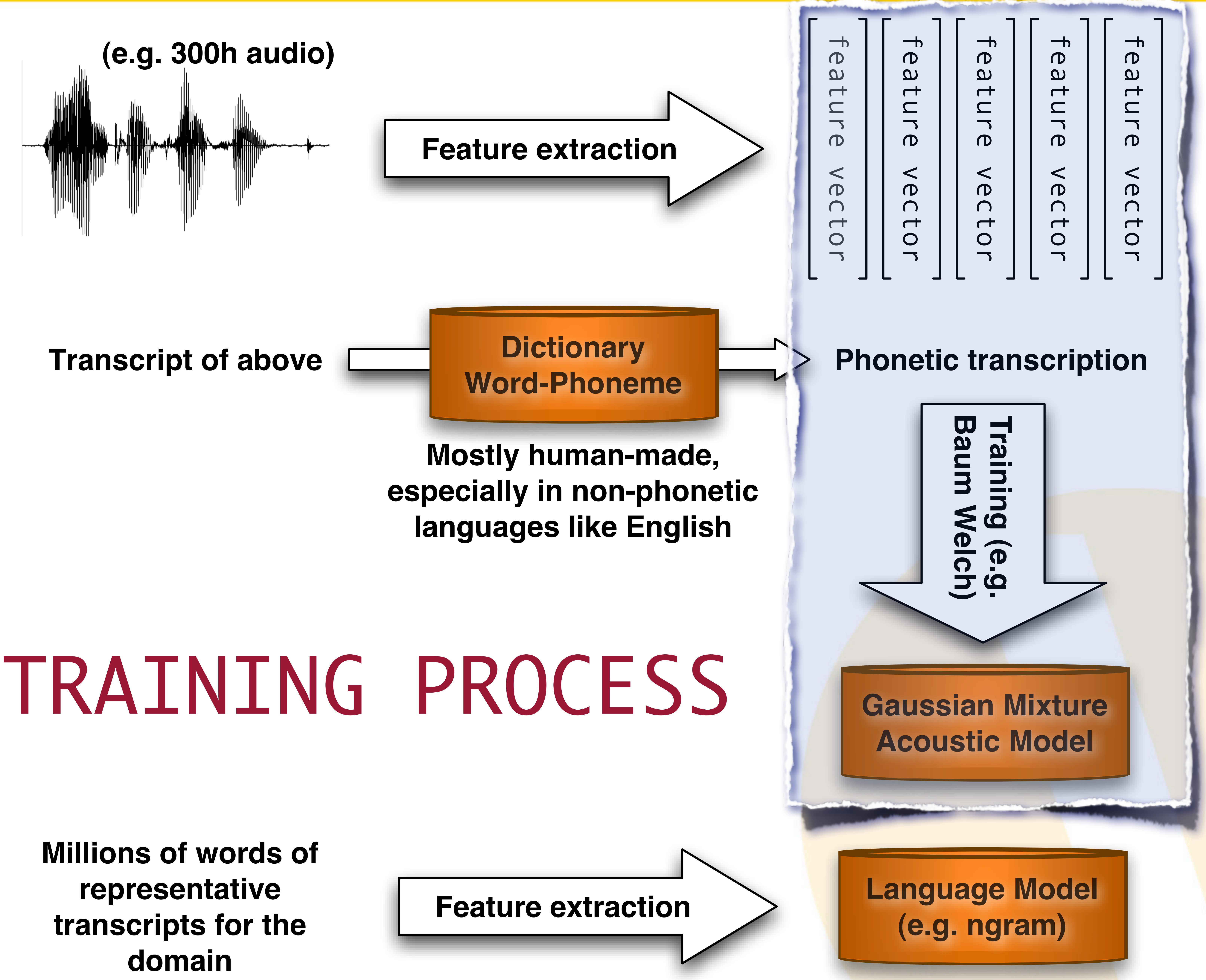
⇒For a phoneme set of 50 phonemes (~English)

potentially 50^3 Triphones
3 states each

- **Reduce space through ‘tying’ states (say down to 10K states)**

- **Every word in the dictionary is represented by a Hidden Markov Model based on these states**

Speech Recognition: Training Process

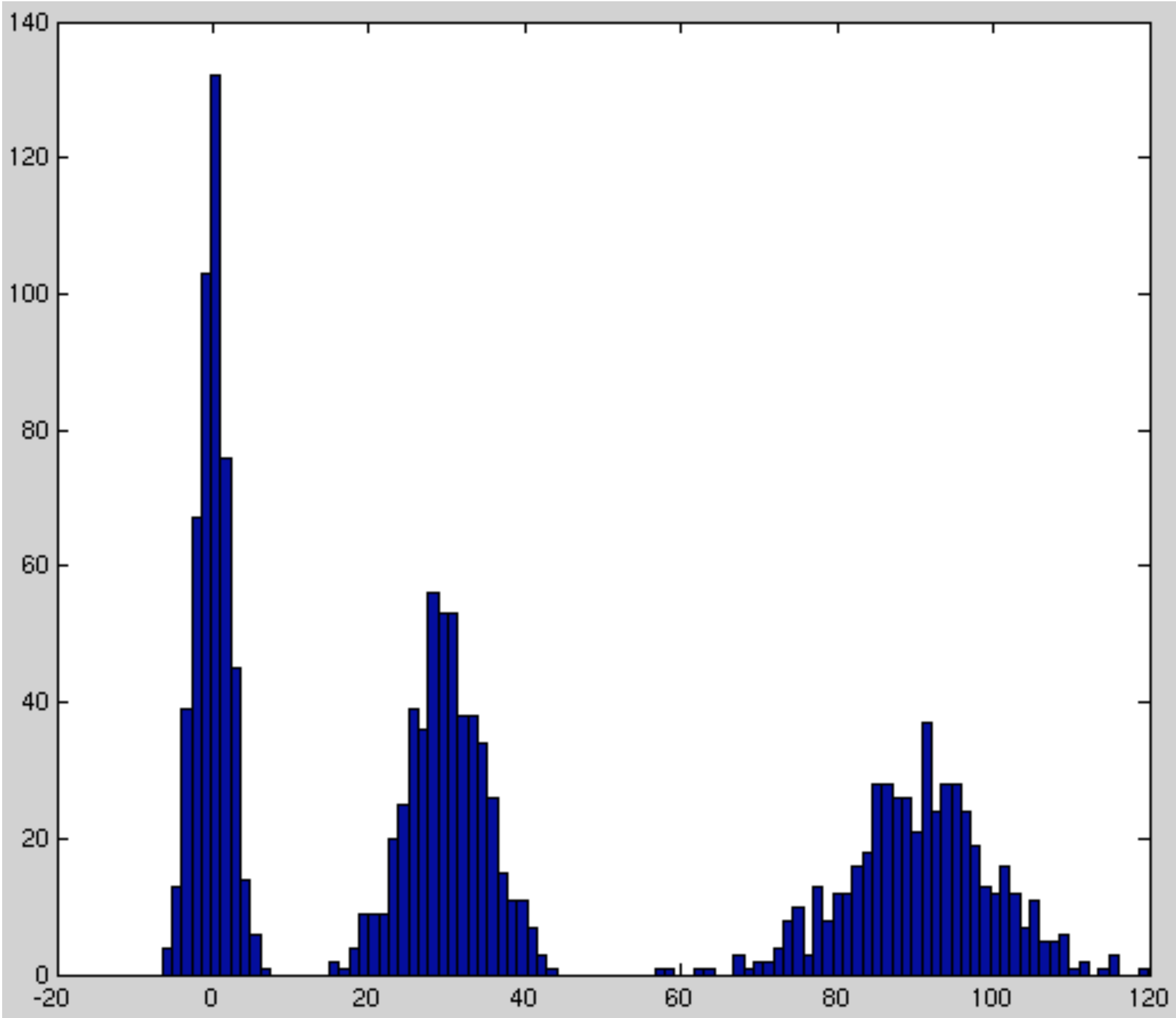


Acoustic model

- **Acoustic model:**
  - Represent the variability for each of these 39 numbers for each state
  - Due to multiple sound instantiations/conditions/speakers/... Gaussian is not a good model.
  - Histogram???
  - Preferred method is a Mixture Gaussian model
  - So in summary:
    - Each phoneme is represented by 3 states
    - Each state is represented by 39 dimensions
    - Each dimension is represented by a mixture Gaussian model (N-means, N-variances, and N-mixture weights -- assuming diagonal cov. matrix)

- **Complexity of Acoustic model in real numbers:**
  - Say 50 phonemes (English)
  - (REAL SYSTEMS) For better accuracy use triphone representation
    - (potentially 50^3 but usually >5K triphones)
  - Each of these has 3 states
  - Each of these has 39 representation dimensions
  - Each dimension has about 32 mixture gaussians
  - 5,000*3*39*(32+32+32) = ~50,000,000 parameters!!
  - (Current SAIL models - 297,000,000 parameters)

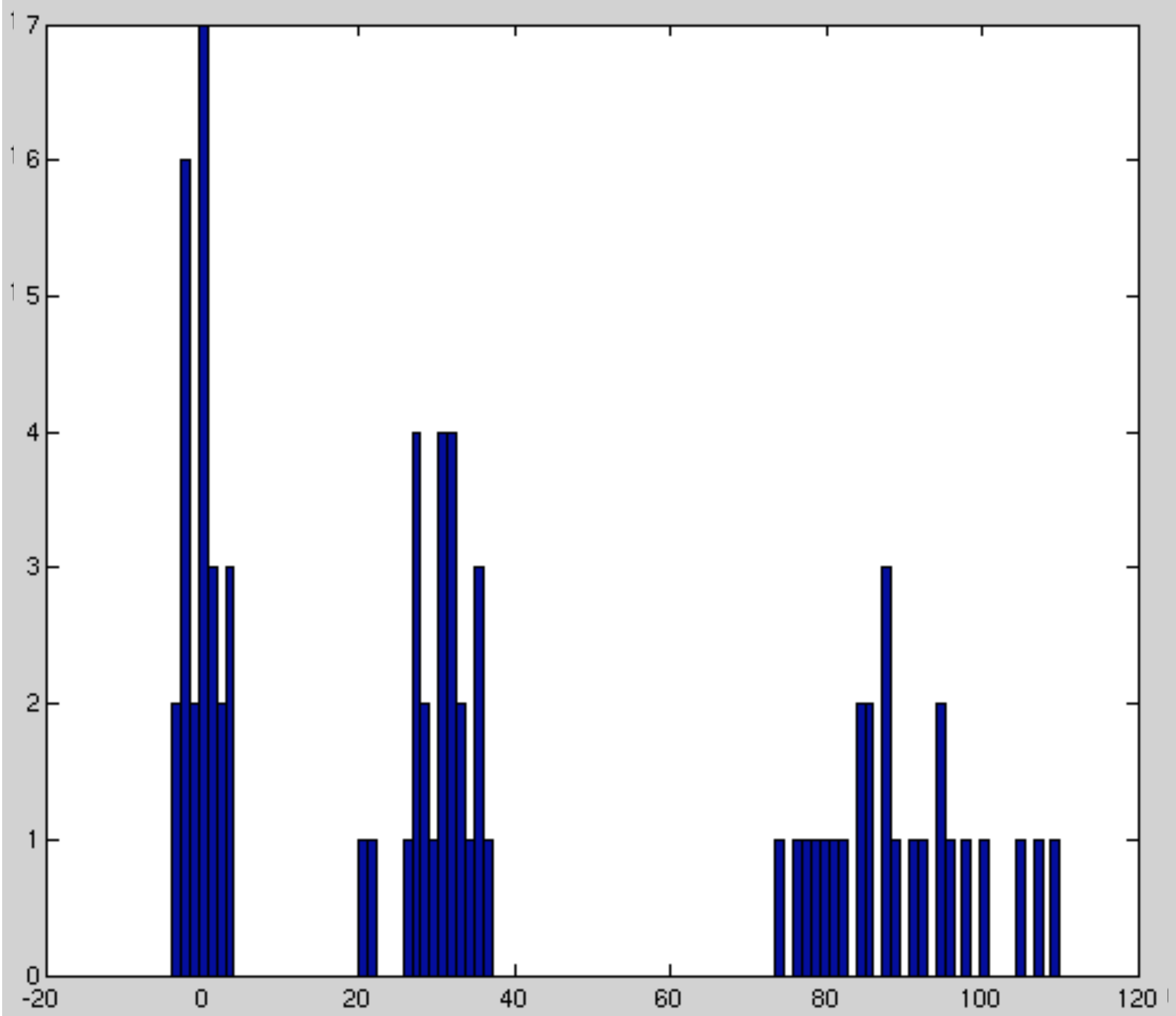
- **Acoustic model:**
  - Represent the variability for each of these 39 numbers for each state
  - Due to multiple sound instantiations/conditions/speakers/... Gaussian is not a good model.
  - Histogram???
  - Preferred method is a Mixture Gaussian model
  - So in summary:
    - Each phoneme is represented by 3 states
    - Each state is represented by 39 dimensions
    - Each dimension is represented by a mixture Gaussian model (N-means, N-variances, and N-mixture weights -- assuming diagonal cov. matrix)



- **Complexity of Acoustic model in real numbers:**
  - Say 50 phonemes (English)
  - (REAL SYSTEMS) For better accuracy use triphone representation
    - (potentially 50^3 but usually >5K triphones)
  - Each of these has 3 states
  - Each of these has 39 representation dimensions
  - Each dimension has about 32 mixture gaussians
  - 5,000*3*39*(32+32+32) = ~50,000,000 parameters!!
  - (Current SAIL models - 297,000,000 parameters)

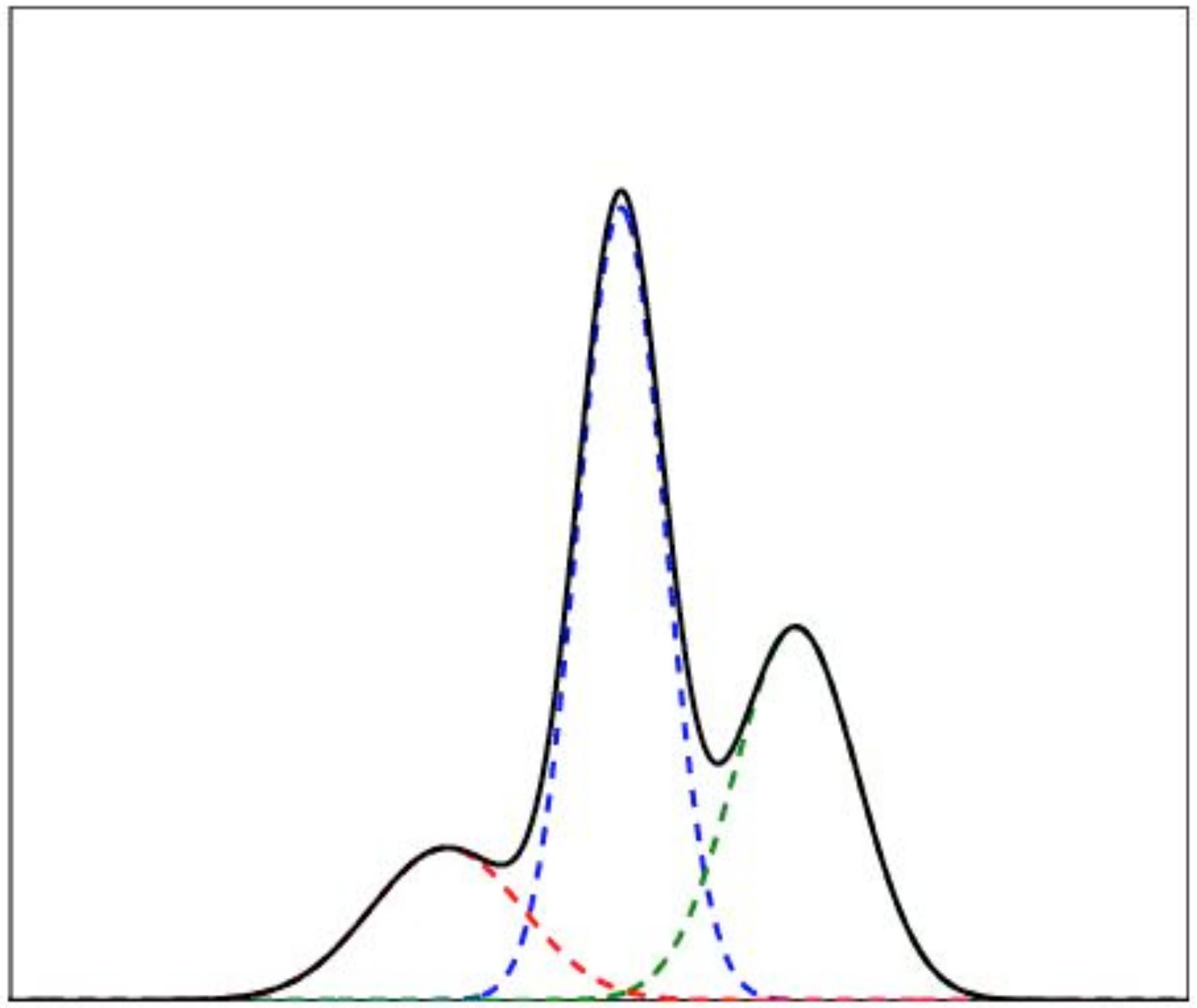
Acoustic model

- **Acoustic model:**
  - Represent the variability for each of these 39 numbers for each state
  - Due to multiple sound instantiations/conditions/speakers/... Gaussian is not a good model.
  - Histogram???
  - Preferred method is a Mixture Gaussian model
  - So in summary:
    - Each phoneme is represented by 3 states
    - Each state is represented by 39 dimensions
    - Each dimension is represented by a mixture Gaussian model (N-means, N-variances, and N-mixture weights -- assuming diagonal cov. matrix)



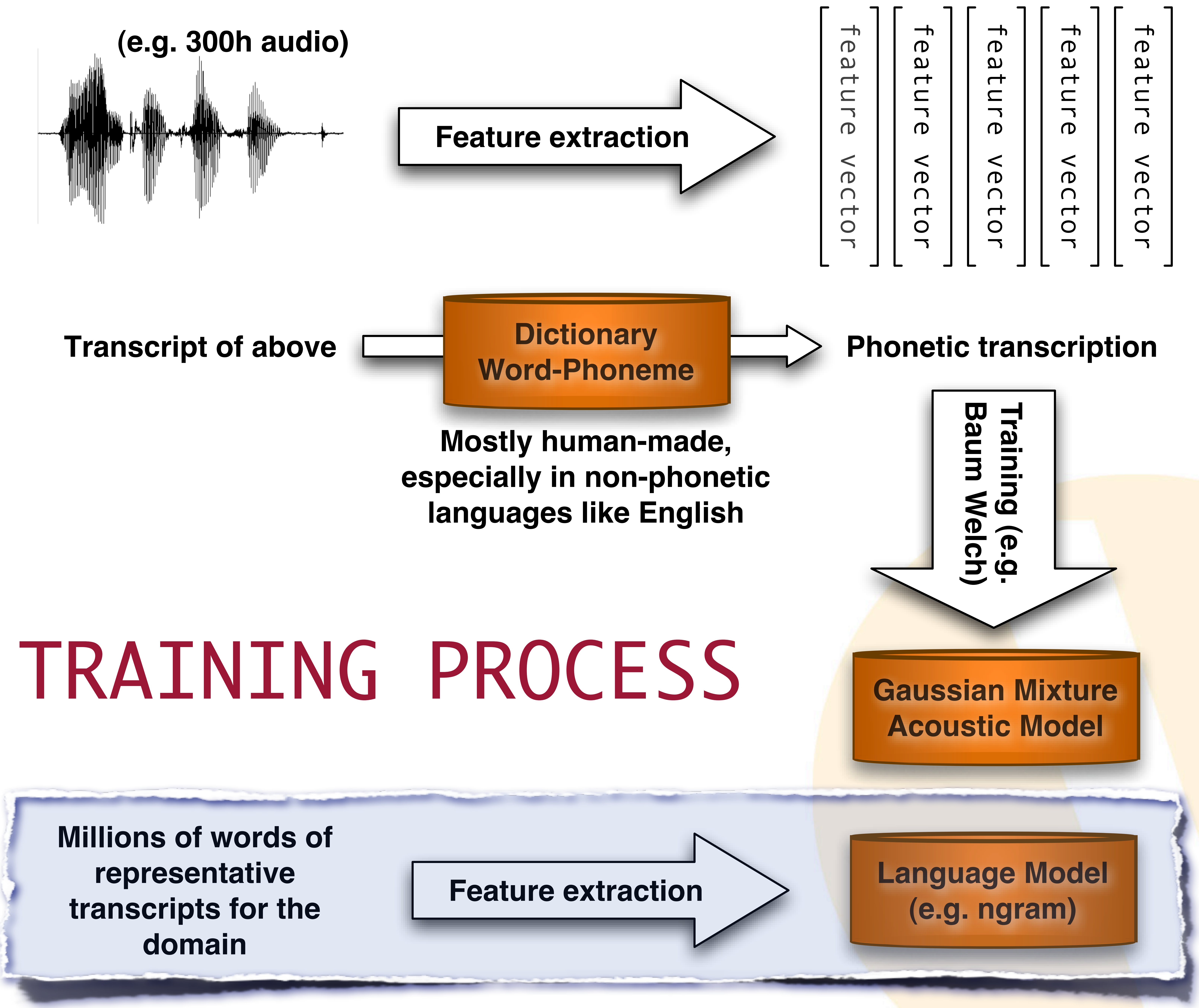
- **Complexity of Acoustic model in real numbers:**
  - Say 50 phonemes (English)
  - (REAL SYSTEMS) For better accuracy use triphone representation
    - (potentially 50^3 but usually >5K triphones)
  - Each of these has 3 states
  - Each of these has 39 representation dimensions
  - Each dimension has about 32 mixture gaussians
  - 5,000*3*39*(32+32+32) = ~50,000,000 parameters!!
  - (Current SAIL models - 297,000,000 parameters)

- **Acoustic model:**
  - Represent the variability for each of these 39 numbers for each state
  - Due to multiple sound instantiations/conditions/speakers/... Gaussian is not a good model.
  - Histogram???
  - Preferred method is a Mixture Gaussian model
  - So in summary:
    - Each phoneme is represented by 3 states
    - Each state is represented by 39 dimensions
    - Each dimension is represented by a mixture Gaussian mode (N-means, N-variances, and N-mixture weights -- assuming diagonal cov. matrix)



- **Complexity of Acoustic model in real numbers:**
  - Say 50 phonemes (English)
  - (REAL SYSTEMS) For better accuracy use triphone representation
    - (potentially 50^3 but usually >5K triphones)
  - Each of these has 3 states
  - Each of these has 39 representation dimensions
  - Each dimension has about 32 mixture gaussians
  - 5,000*3*39*(32+32+32) = ~50,000,000 parameters!!
  - (Current SAIL models - 297,000,000 parameters)

Speech Recognition: Training Process



- **Second term:**

$$\hat{W} = \arg\max_{W \in D} \underbrace{P(O|W)}_{\text{Acoustic Model}} \underbrace{P(W)}_{\text{Language Model}}$$

- **P(W) can be extracted from existing text:**

$$P(W) = P(W_1, W_2, ...W_n) = P(W_1)P(W_2|W_1)P(W_3|W_1W_2....)P(W_n|W_1W_2...W_{n-1})$$

- **For simplicity and feasibility aproximate with:**

$$P(W) = P(W_1, W_2, ...W_n) = P(W_1)....P(W_{n-1}|W_{n-3}W_{n-2})P(W_n|W_{n-2}W_{n-1})$$

- **When we don't have enough data - next best:**
  - $p(w_3|w_1, w_2) =$

$$\begin{array}{rl} \text{if(trigram exists)} & P_3(w_1, w_2, w_3) \\ \text{else if(bigram } w_1, w_2 \text{ exists)} & BOW(w_1, w_2)P(w_3|w_2) \\ \text{else} & P(w_3|w_2) \end{array}$$

USC

SAiL

Virtual Character Complications

- **Learn from large amounts of existing text**
- **Dealing with data sparsity:**
  - Smoothing
  - Background models
  - Mining
  - etc

One 'UNIVERSITY' unigram:

```
-3.86769  UNIVERSITY  -0.5197889
```

Results in 1056 bigrams

```
-3.120121 UNIVERSITY WORK -0.07356837
```

and 1650 trigrams

```
-1.634784 HIS UNIVERSITY WORK
```

Virtual character data: Really data starved. Very few potential n-grams seen, especially 2+grams	Background LM on the same data. Much better coverage but not of this domain.	Smoothing w/background covers the language possibilities better, but the probabilities are 'flat'
\data\ ngram 1=1422 ngram 2=6613 ngram 3=9943	\data\ ngram 1=1422 ngram 2=370422 ngram 3=2231793	\data\ ngram 1=5353 ngram 2=2650680 ngram 3=6881435

Speech Recognition: Recognition (Decoding) Process

- **Decoding:**

$$\hat{W} = \arg\max_{W \in D} P(O|W)P(W)^N$$

- **Every frame:**
  - Birth of new words: this is probabilistic so hundreds of words are potentially starting every 10ms
  - Lexical Tree like search makes this faster (i.e. If we have seen phonemes X Y then all the words starting from X Y will be searched, but not remaining words)
  - As we move forward we can prune paths based on:
    - Maximum total alive words at any time instant
    - Maximum new words at any time instant
    - Pruning low probability paths by deeming them un-viable
    - Constraining total search space (dangerous), etc
  - Pruning reduces performance, so a good LM, and AM reduces the probability of pruning good paths

- **Real time systems, “bad” LM, large/mismatched domains**

Summary

- **ASR aspects**
  - Needed:
    - Representative audio
    - Transcriptions of the audio
    - Good HMM models (word -> phoneme dictionaries) for all transcripts
    - Large amounts of representative text (in the millions)
- **Other real-system complications:**
  - Click to talk: needed to reduce search space and ambiguity
  - Without it we need:
    - VAD: Voice Activity Detection can do the coarse segmentation of speech--non-speech
    - Utterance segmentation: needed for breaking up continuous streams of audio (e.g. this presentation)
    - If both absent: ASR is near useless.

  - Speed
  - Audio quality

- Acoustic models:
  - A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Bilmes, J.A., International Computer Science Institute, Vol. 4, 1998
    http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.4498
  - A tutorial on hidden Markov models and selected applications inspeech recognition, LR Rabiner Proceedings of the IEEE, Vol. 77, No. 2. (1989), pp. 257-286.
    http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=18626&isnumber=698
  - Abhinav Sethy, Panayiotis Georgiou, Bhuvana Ramabhadran, and Shrikanth Narayanan. An iterative relative entropy minimization based data selection approach for n-gram model adaptation. IEEE Transactions on Speech, Audio and Language Processing, In press, 2008.