Acoustic Modeling
Variability in the Speech Signal
Environmental Robustness

Kjell Elenius

Speech, Music and Hearing
KTH

Ch 9 Acoustic Modeling

- Variability in the Speech Signal
- How to Measure Speech Recognition Errors
- Signal Processing – Extracting Features
- Phonetic Modeling – Selecting Appropriate Units
- Acoustic Modeling – Scoring Acoustic Features

Ch 9 Acoustic Modeling 1(4)

- Variability in the Speech Signal
  - Context Variability
  - Style Variability
  - Speaker Variability
  - Environment Variability
- (How to Measure Speech Recognition Errors)
- Signal Processing – Extracting Features
  - Signal acquisition
  - End-Point Detection
  - MFCC and Its Dynamic Features
  - Feature Transformation

Ch 9 Acoustic Modeling 2(4)

- Phonetic Modeling – Selecting Appropriate Units
  - Comparison of Different Units
  - Context Dependency
  - Clustered Acoustic-Phonetic Units
  - Lexical Baseforms
- Acoustic Modeling – Scoring Acoustic Features
  - Choice of HMM Output Distributions
  - Isolated vs. Continuous Speech Training

Ch 9 Acoustic Modeling 3(4)

- Adaptive Techniques – Minimizing Mismatches
  - Maximum a Posteriori (MAP)
  - Maximum Likelihood Linear Regression (MLLR)
  - MLLR and MAP Comparison
  - Clustered Models
- Confidence Measures: Measuring the Reliability
  - Filler Models
  - Transformation Models
  - Combination Models

Ch 9 Acoustic Modeling 4(4)

- Other Techniques
  - Neural Networks
  - Segment Models
    - Parametric Trajectory Models
    - Unified Frame- and Segment-Based Models
  - Articulatory Inspired Modeling
    - HMM2, feature asynchrony, multi-stream (separate papers)
  - Use of prosody and duration

Acoustic model requirements

- Goal of speech recognition
  - Find word sequence with maximum posterior probability

$$\hat{W} = \arg\max_{w} P(\mathbf{W}|\mathbf{X}) = \arg\max_{w} \frac{P(\mathbf{W})P(\mathbf{X}|\mathbf{W})}{P(\mathbf{X})} \propto \arg\max_{w} P(\mathbf{W})P(\mathbf{X}|\mathbf{W})$$

  - One linguistic P(**W**) and one acoustic model P(**X**|**W**)
  - In large vocabulary recognition, phonetic modeling is better than word modeling
    - Training data size
    - Tying between similar parts of words
    - Recognition speed
  - The acoustic model should include
    - variation due to speaker, pronunciation, environment, coarticulation
    - dynamic adaptation

9.1 Variability in the Speech Signal

- Context
  - Linguistic
    - homonyms: same pronunciation but meaning dependent on word context
  - Acoustic
    - coarticulation, reduction effects
- Speaking style
  - isolated words, read-aloud speech, conversational speech
- Speaker
  - dependent, independent, adaptive
- Environment
  - background noise, reverberation, transmission channel

9.2 How to Measure Speech Recognition Errors

- Dynamic programming to align recognised and correct strings
- Gives optimistic performance
- Discards phonetic similarity

$$\text{Word error rate} = 100\% * \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{No. of words in the correct sentence}}$$

9.3 Signal Processing – Extracting Features

- Purpose
  - Reduce the data rate, remove noise, extract useful features
- Signal Acquisition
- End-Point Detection
- MFCC and its Dynamic Features
- Feature Transformation

9.3.1 Signal acquisition

Sampling rate	Relative Error-rate Reduction
8 kHz	Baseline
11 kHz	+10%
16 kHz	+10%
22 kHz	+0%

Effect of sampling rate on the performance

- Practical consideration on slow machines: buffering
- Children's speech benefit from higher sampling rate

9.3.2 End-Point Detection

- Two-class pattern classifier selects intervals to be recognised
- Based on energy, spectral balance, duration
- Exact end-point positioning not critical
  - Low rejection rate more important than low false acceptance
  - Lost speech segments cause errors, accepted external noise can be rescued by the recogniser
- Adaptive algorithm (EM) better than fixed threshold
- Buffering necessary

9.3.3 MFCC and Its Dynamic Features

- Temporal changes important for human perception
- *Delta coefficients:* 1st and 2nd order time derivative
  - Capture short-time dependencies
- Typical state-of-the-art system
  - 13th order MFCC c_k
  - 13th-order 40 ms 1st order deltas $\Delta c_k = c_{k+2} - c_{k-2}$
  - 13th-order 2nd order deltas $\Delta\Delta c_k = \Delta c_{k+1} - \Delta c_{k-1}$
- Often computed as regression lines

Feature set	**Rel. Error Reduction**
13^{th}-order LPCC	Baseline
13^{th}-order MFCC	+10%
16^{th}-order MFCC	+0%
$+1^{st}$ and 2^{nd} order deltas	+20%
$+3^{rd}$ order deltas	+0%

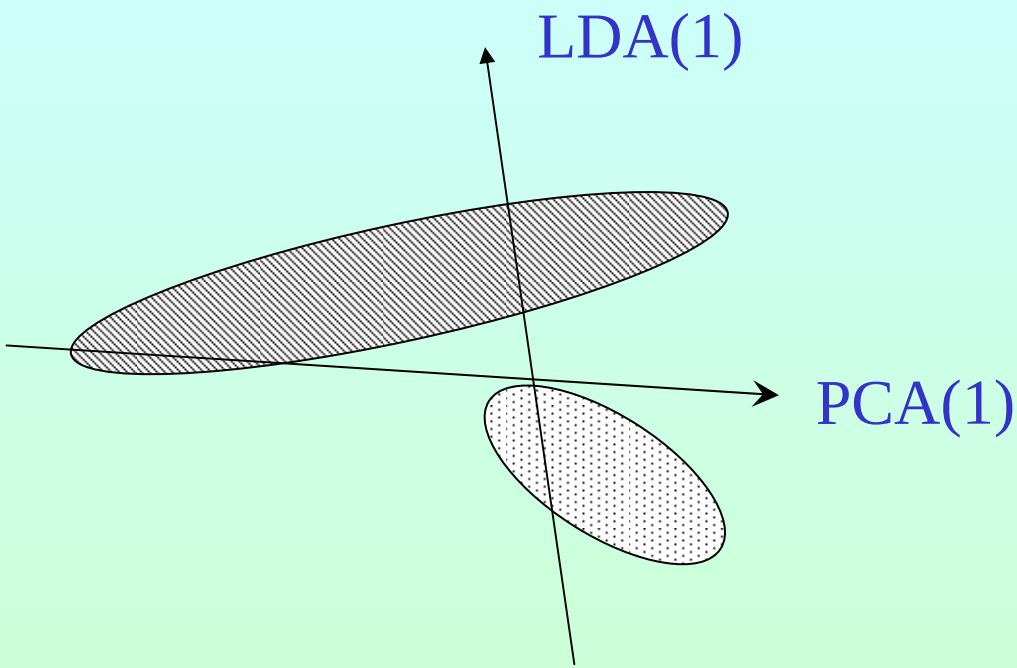
9.3.4 Feature Transformation: PCA

- Principal-Component Analysis (PCA)
  - Also known as Karhunen-Loewe transform
  - Maps a large feature vector into smaller dimensional vector
  - New basis vectors: eigenvectors, ordered by the amount of variability they represent (eigenvalues)
  - Discard those with the smallest eigenvalues
  - The transformed vector elements are uncorrelated

9.3.4 Feature Transformation: LDA

- LDA: Linear Discriminant Analysis
- Transform the feature vector into a space with *maximum class discrimination*
- Method
  - "Quotient" between *Between Class Scatter* and *Within Class Scatter*
  - The eigenvectors of this matrix constitute the new dimensions
  - The first LDA eigenvectors represent the directions in which the *class discrimination* is maximum
    - PCA eigenvectors represent directions with class independent variability

PCA vs LDA



PCA finds directions with maximum class-independent variability
LDA finds directions with maximum class discrimination

9.3.4 Feature Transformation:
Frequency warping for vocal tract length normalisation

- Linear or piece-wise linear scaling of the frequency axis to account for varying vocal tract size
  - Shift of center frequencies of the mel-scale filter bank
  - Scaling of center frequencies of linear frequency filter bank
  - In theory, phoneme dependent scaling is necessary
  - Phoneme-independent scaling used in practice, works reasonably well.
    - 10% relative error reduction among adult speakers
    - Larger reduction when children use adult phone models

9.4 Phonetic Modeling – Selecting Appropriate Units

- What is the best base unit for a continuous speech recogniser?
- Possible units
  - Phrase, word, syllable, phoneme, allophone, subphone
- Requirements
  - Accurate
    - Can be recognised with high accuracy
  - Trainable
    - Can be well trained with the given size of the training data
  - Generalizable
    - Words not in the training data should be modelled with high precision

9.4.1 Comparison of Different Units

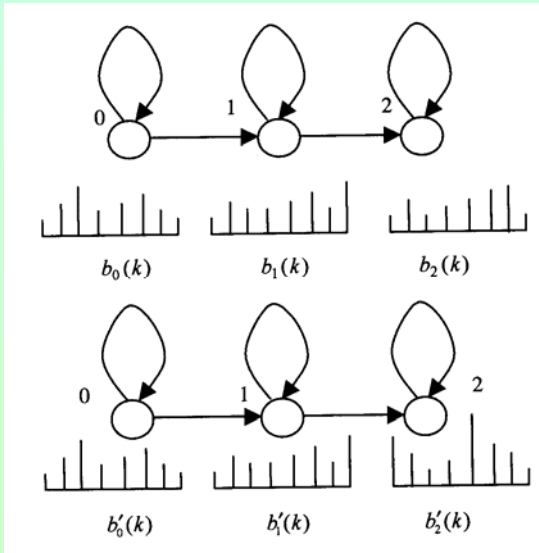
- Phrase
  - + Captures coarticulation for a whole phrase
  - – Very large number. Common phrases might be trainable
- Word
  - + Intra-word, but not inter-word coarticulation is captured
    - Requires word-pair training
  - – Very large number, large vocabulary training unrealistic
- Syllable
  - + Close tying with prosody (stress, rhythm)
  - – Coarticulation at endpoints not captured, Large number
- Phone
  - + Low number (around 50)
  - – Very sensitive to coarticulation
- Context-dependent phone (triphone, diphone, monophone)
  - + Captures coarticulation from adjacent phones
  - – High number of triphones (125 000)

9.4.2 Context Dependency

- Triphones cover the dependence from immediately neighboring phonemes
- Dependence not captured:
  - Certain coarticulation
    - Phones at longer distance (e.g., lip rounded, retroflex, nasal)
    - Across word boundaries (often)
  - Stress information (normally)
    - Lexical stress (**im**port vs. im**port**)
    - Sentence-level stress
    - Contrastive stress
    - Emphatic stress

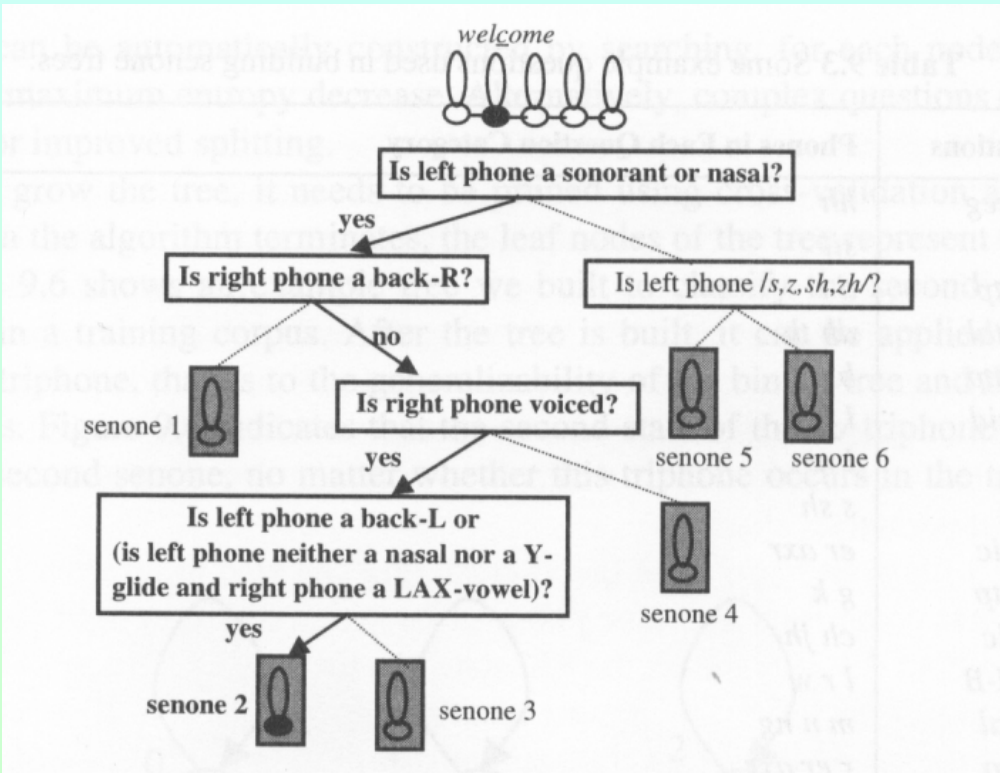
9.4.3 Clustered Acoustic-Phonetic Units

- Parts of certain context-dependent phones are similar
  - The subphone state can be a basic speech unit
  - The very large number of states is reduced by clustering (tying)
  - *Senones*
  - State-based clustering can keep dissimilar states of two phone models apart but merge the similar ones
  - Better parameter sharing than in phone-based tying
  - The first two states can be tied:



Predict Unseen Triphones

- Which senones to represent a triphone that does not exist in the training data?
- Decision tree



Decision tree for selecting senone for 2nd state of /k/ triphone

Unit Performance Comparison

Units	Rel. Error Reduction
Context-independent phone	Baseline
Context-dependent phone	+25%
Clustered triphone	+15%
Senone	+24%

Relative error reduction for different modelling units. The reduction is relative to the preceding row

9.4.4 Lexical Baseforms

- Dictionary contains standard pronunciation
  - Need alternative pronunciations
- Phonological rules to modify word boundaries and to model reduced speech
- Proper names often not included in dictionaries
  - Need to be derived automatically
  - Rule-based letter-to-sound conversion not good for English
  - Need trainable LTS converter
  - Neural networks, HMM, CART

CART-based LTS Conversion

- Questions in a context window, size around 10 letters
- Give more weight to nearby context
  - Example: “Is the second letter to the right ‘p’?”
  - Use a transcribed dictionary for generating the tree
  - Splitting criterion: Entropy reduction
  - Conversion error 8% on English newspaper text
  - Error types
    - Proper nouns and foreign words
    - Generalisation
  - Exception dictionary necessary

Pronunciation Variability

- Multiple entries in dictionary or finite state machine
- Modest error reduction (5-10%) by current approaches
  - Allows too much variability
- Studies indicate high potential

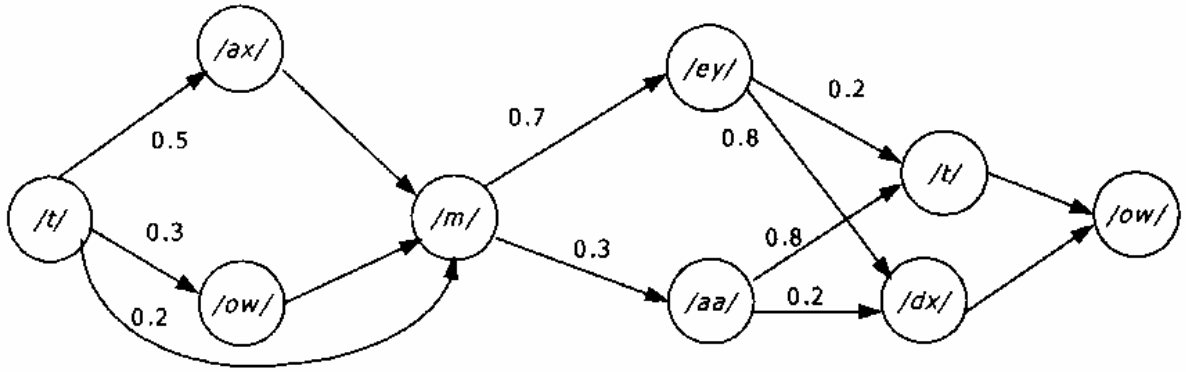


Figure 9.7 A possible pronunciation network for word *tomato*. The vowel */ey/* is more likely to flap, thereby having a higher transition probability into */dx/*.

Pronunciation Variability: Possible Research Directions

- Simulations indicate possible error reduction
  - Factor 5-10 (McAllaster et al, 1998)
- Experiments not as successful
  - Possibly 35% relative (Yang et al, 2002)
  - In practice, 5-10%
- Why no improvement?
  - Gaussian mixtures can model phone insertion and substitution
    - Rules for phone deletion still of value (Jurafski et al, 2001)
  - Rules tend to over-generate, allow too much variability
    - Need to be specific for each speaker (style, accent, etc.)
    - Inter-rule dependence

9.5 Acoustic Modeling – Scoring Acoustic Features

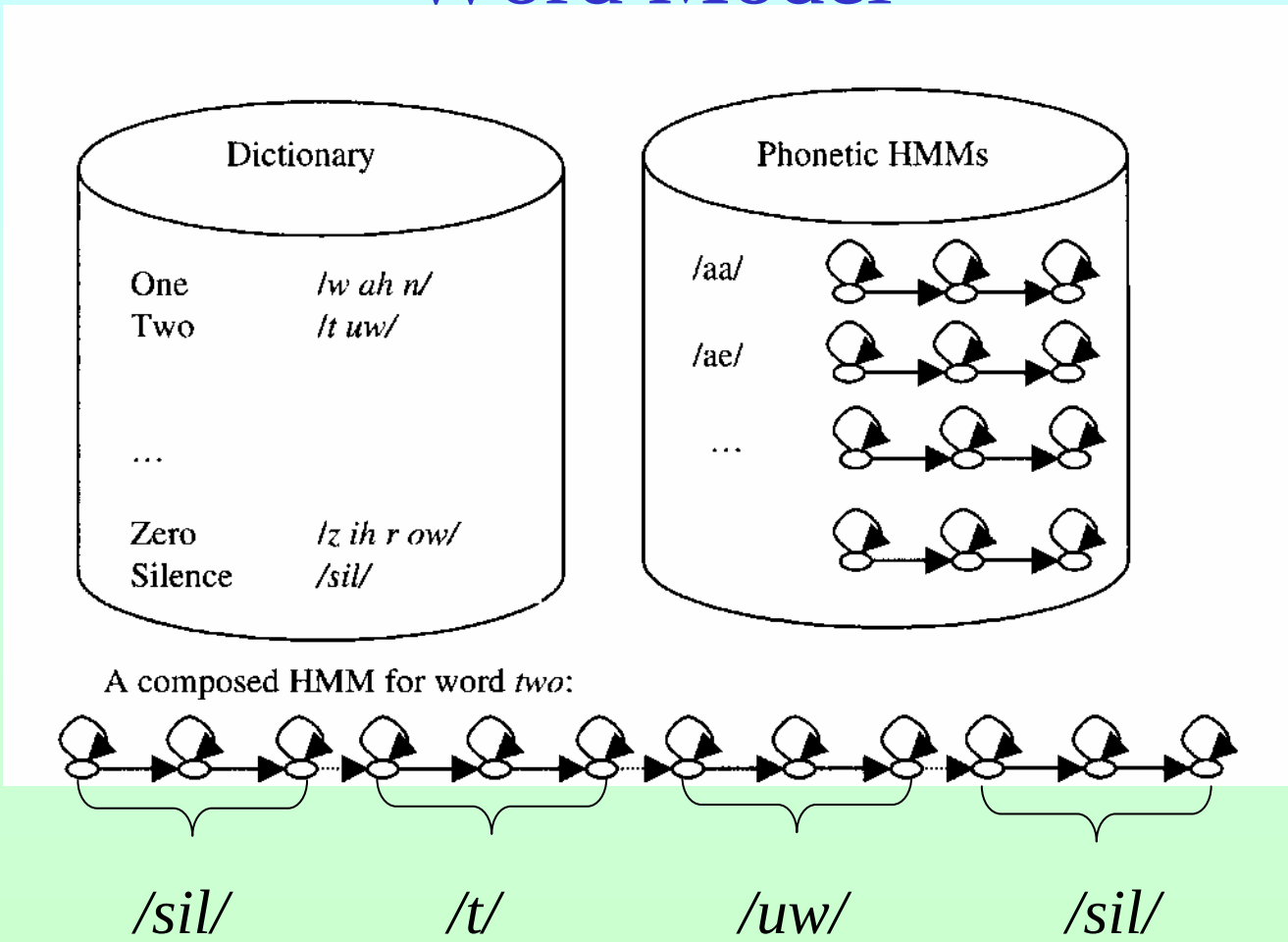
Choice of HMM Output Distributions

- Discrete, continuous, or semicontinuous HMM?
  - If training data is small, use DHMM or SCHMM
  - Multiple codebooks
    - E.g. separate codebooks for static, delta and acceleration features
  - Number of mixture components
    - With sufficient training data, 20 components reduce SCHMM error by 15-20%

Isolated vs. Continuous Speech Training

- In isolated word speech recognition, each word is trained in isolation
  - Straight-forward Baum-Welch training
- In continuous and phoneme-based speech recognition, each unit is trained in varying context
- Phones and words are connected by null transitions

Concatenation of Phone Models into a Word Model



Composite Sentence HMM

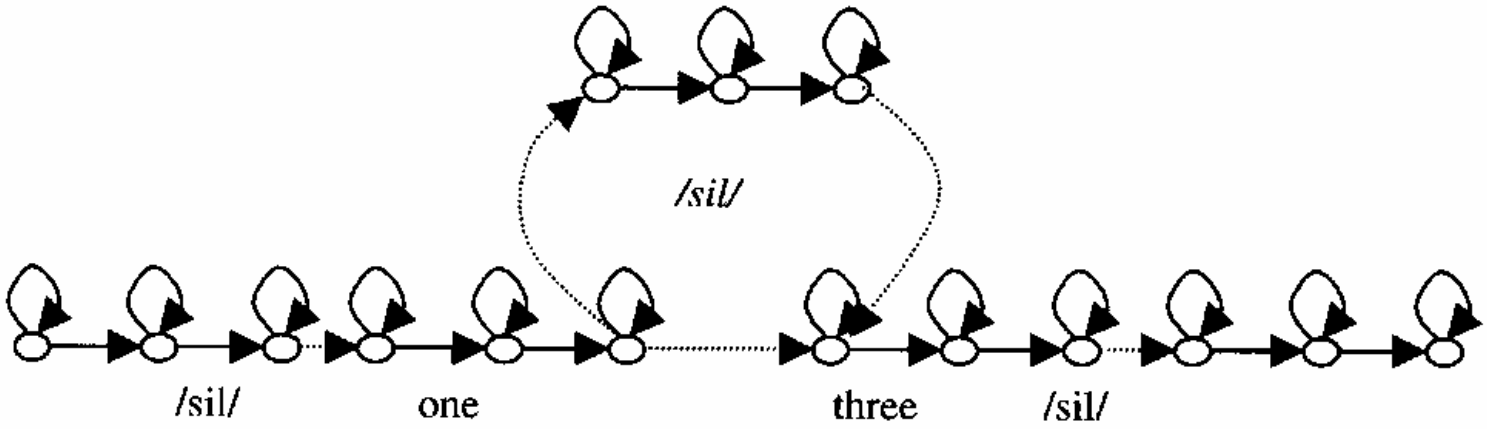


Figure 9.10 A composite sentence HMM. Each word can be a word HMM or a composite phonetic word HMM, as illustrated in Figure 9.9.

9.7 Confidence Measures

- The system's belief in its own decision
- Important for
  - out-of-vocabulary detection
  - repair probable recognition errors
  - word spotting
  - training
  - unsupervised adaptation
- Theory $$P(\mathbf{W}\mid\mathbf{X}) = \frac{P(\mathbf{W})P(\mathbf{X}\mid\mathbf{W})}{P(\mathbf{X})} = \frac{P(\mathbf{W})P(\mathbf{X}\mid\mathbf{W})}{\sum_{\mathbf{W}} P(\mathbf{W})P(\mathbf{X}\mid\mathbf{W})}$$
- Good confidence estimator if the denominator is not ignored

9.7.1 Filler Models

- Represent the denominator P(**X**) by a general-purpose recognizer
  - E.g. phoneme recognizer
- Run the two recognizers in parallel
- Individual word confidence is derived by accumulating the ratio over the duration of a recognised word

9.7.2 Transformation Models

- Idea
  - Some phonemes may be more important for the confidence score
  - Give more weight to these

$$\wp_i(x) = ax + b$$

  - The confidence of phoneme *i* is transformed
  - Word confidence

$$CS(w) = \sum_{i=1}^{N} \wp_i(x_i) / N$$

Phoneme Specific Confidence Weights

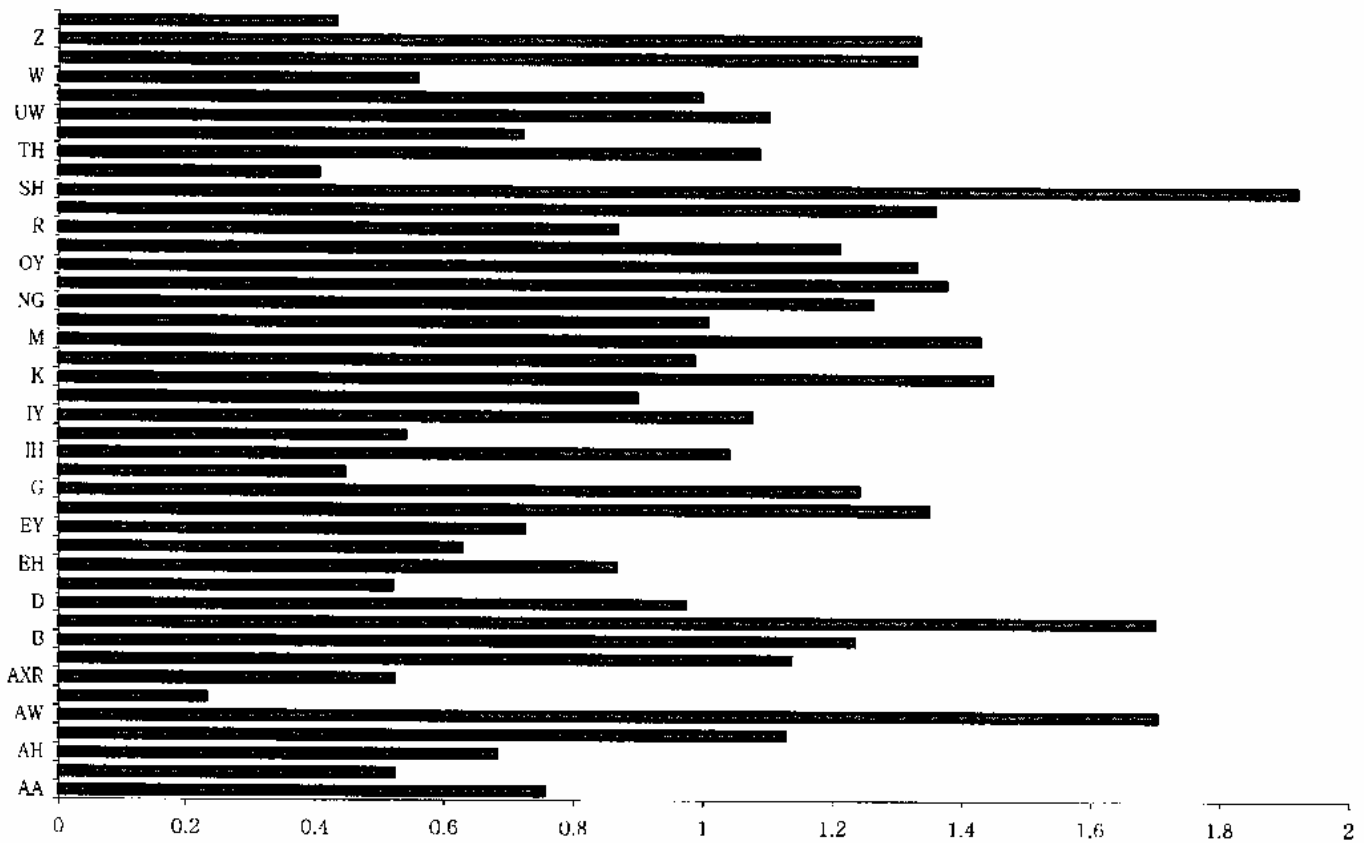


Figure 9.12 Transformation parameter *a* for each context-independent phone class. The weight of consonants is typically larger than that of vowels [63].

Confidence Accuracy Improvement by Transformation Model

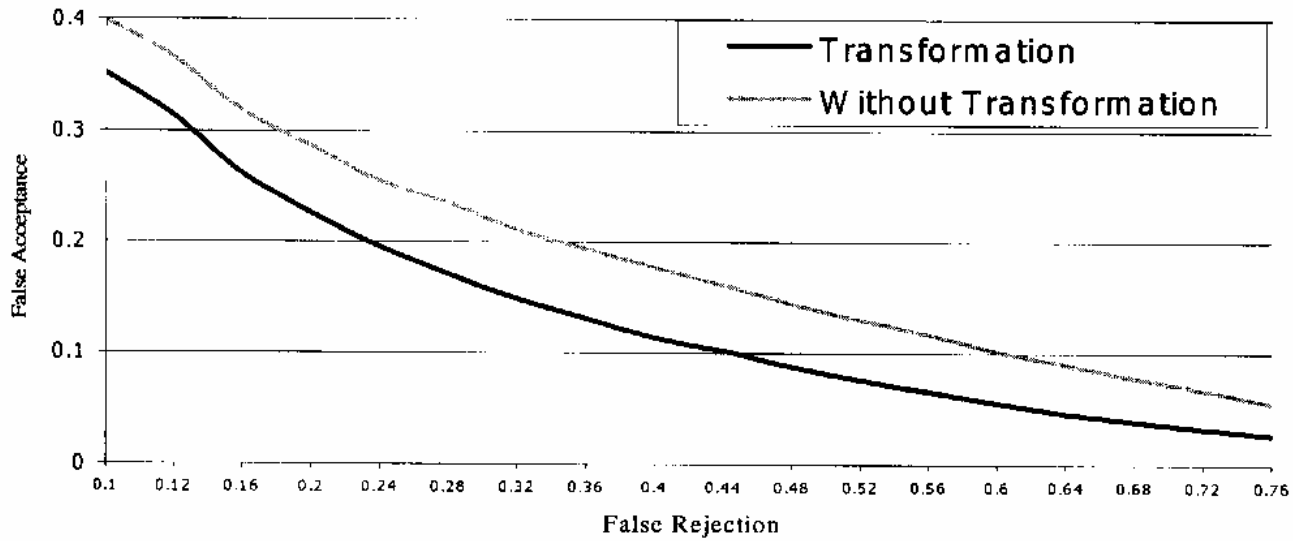


Figure 9.13 The ROC curve of phonetic filler models with and without optimal feature transformation [63].

9.7.3 Combination Models

- Combine different features to a confidence measure
  - Word stability using different language models
  - Average number active words at end of utterance
  - Normalized acoustic score per frame in word
- Combination metric is insignificant
  - linear classifier works well

9.8 Other Techniques

- In addition to HMM
  - Neural Networks
  - Segment Models
  - 2D HMM
  - Bayesian networks
  - Multi-stream
  - Articulatory oriented representation
  - Prosody and duration
  - Long range dependencies

9.8.1 Artificial Neural Networks (ANN)

- Good performance for phoneme classification and isolated, small-vocabulary recognition
- Problem
  - Basic neural nets have trouble handling patterns with timing variability (such as speech)
    - Alignment, training, decoding
- Approaches
  - Recurrent neural networks
    - Memory of previous outputs or internal states
  - Time Delay Neural Networks
    - A time sequence of acoustic features are input to the net
  - Integration with HMM (Hybrid system)
    - The ANN replaces the Gaussian mixture densities

Time Delay Neural Network (TDNN)

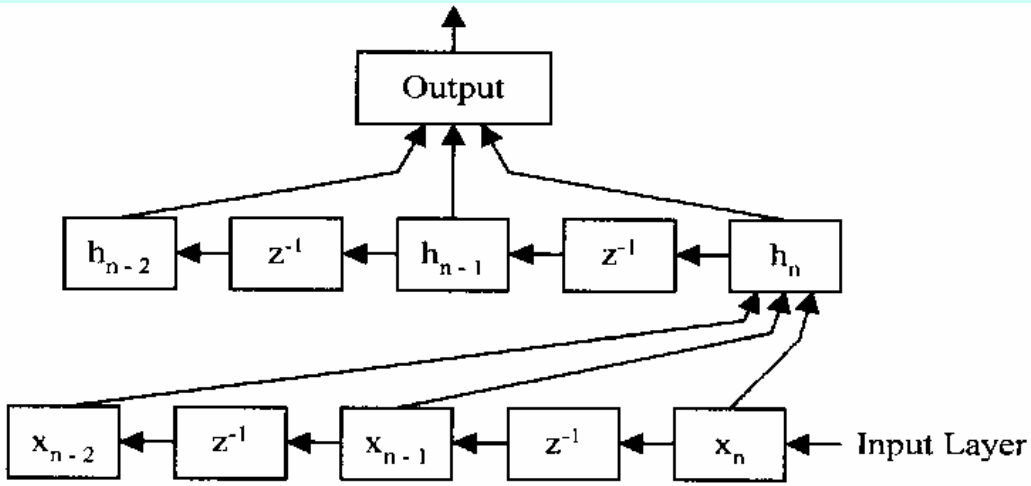


Figure 9.15 A time-delay neural network (TDNN), where the box h_t denotes the hidden vector at time t, the box x_t denotes the input vector at time t, and the box z^{-1} denotes a delay of one sample.

Recurrent Network

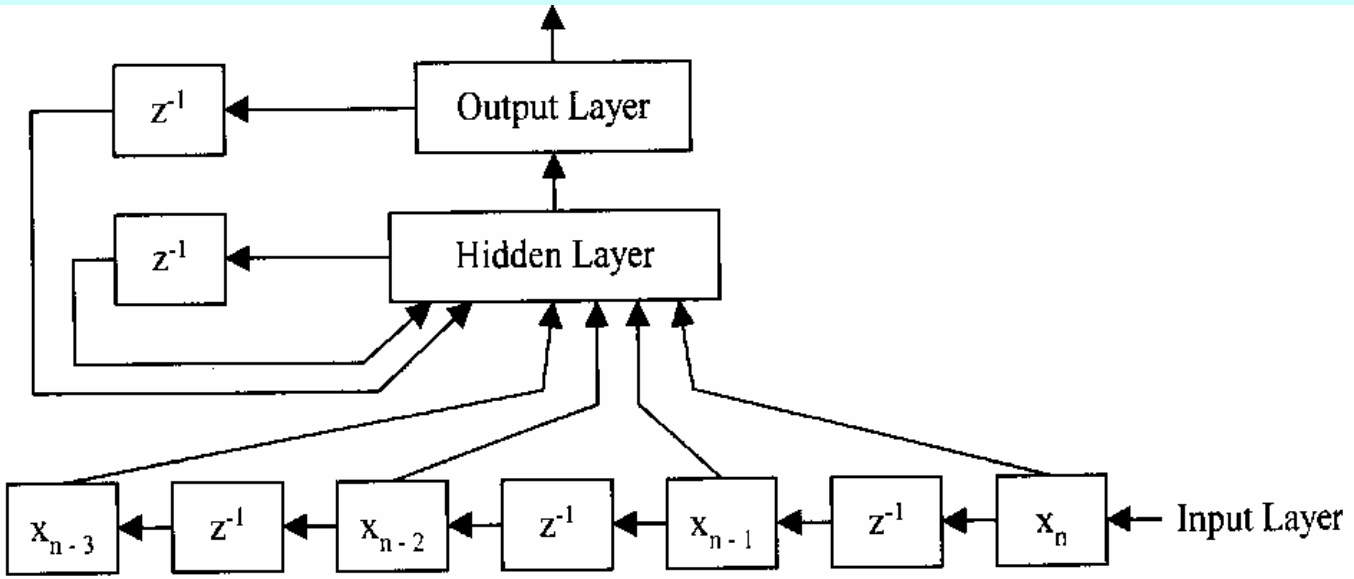
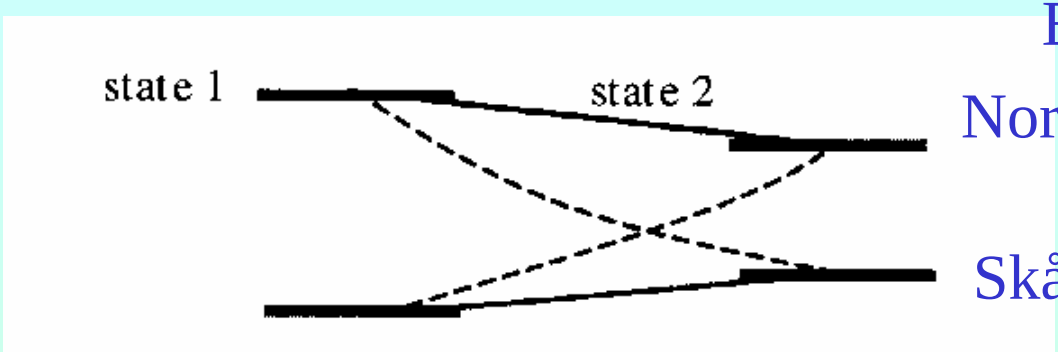


Figure 9.14 A recurrent network with contextual inputs, hidden vector feedback, and output vector feedback.

9.8.2 Segment Models

- Problem
  - The HMM output-independence assumption results in a stationary process (constant mean and variance) in each state
    - Bad model, speech is non-stationary
    - Delta and acceleration features help, but the problem remains
  - Phantom trajectories can occur
    - Trajectories that did not exist in the training data
- Approach
  - An interval trajectory rather than a single frame value is matched
  - Parametric Trajectory Models
  - Unified Frame- and Segment-Based Models
  - Heavily increased computational complexity

Phantom Trajectories



Example:

Norrländsk accent

Skånsk accent

Mixture component sequences that never occurred in the same utterance during training, are allowed during recognition

Standard HMM allows every frame in an utterance to come from a different speaker

Parametric Trajectory Models

- Model a speech segment with curve-fitting parameters
  - Time-varying mean
  - Linear division of the segment in constant number of samples
  - Multiple mixtures possible
  - Low number of trajectories needed for speaker-independent recognition
    - Seems to help the phantom trajectory problem
  - Estimation by EM algorithm
  - Modest improvement over HMM

Unified Frame- and Segment-Based Models

- HMM and segment model (SM) approaches are complementary
  - HMM: detailed modeling but quasi-stationary
  - SM: models transitions and longer-range dynamics but coarse detail
- Combine HMM and SM

$$p(\mathbf{X} \mid \text{Unified model}) = p(\mathbf{X} \mid \text{HMM})\, p(\mathbf{X} \mid \text{SM})^{a}$$

- 8% WER reduction compared to HMM Whisper
- (Method developed by course book co-author)

Research Progress Evolution

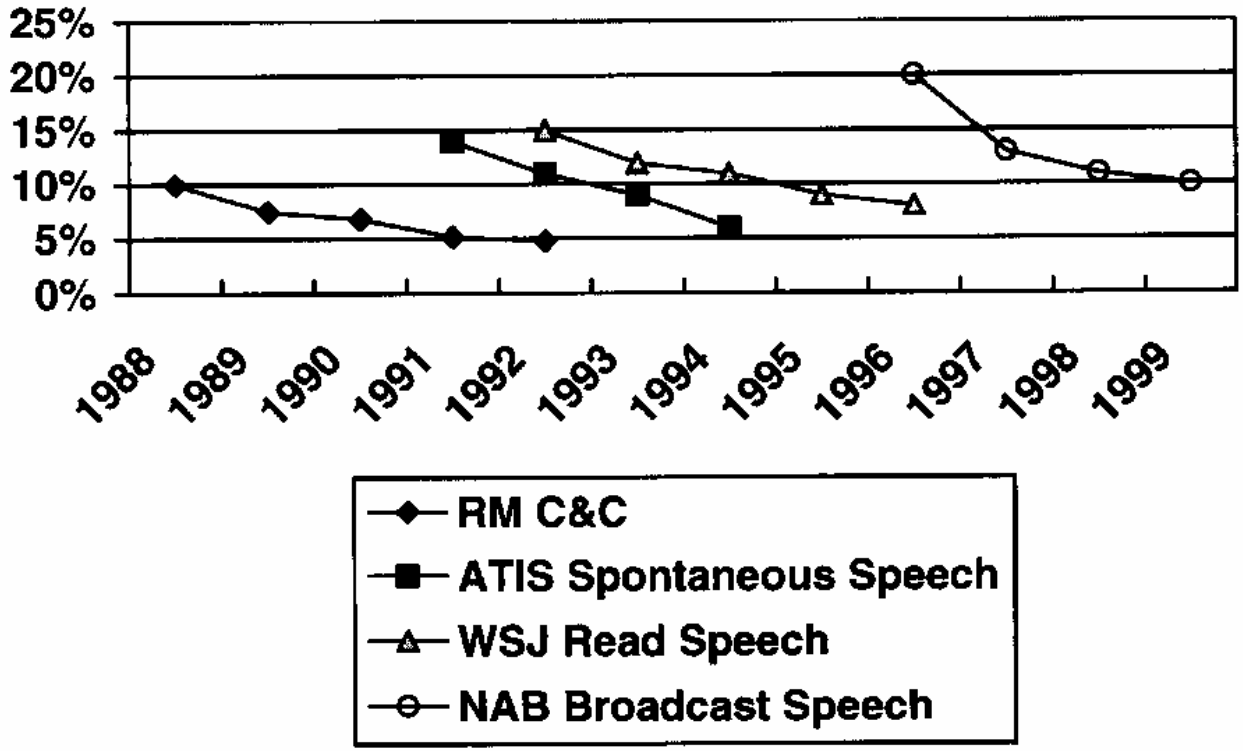


Figure 9.18 History of DARPA speech recognition word-error-rate benchmark evaluation results from 1988 to 1999. There are four major tasks: the Resource Management command and control task (RM C&C, 1000 words), the Air Travel Information System spontaneous speech understanding task (ATIS, 2000 words), the *Wall Street Journal* dictation task (WSJ, 20,000 words), and the Broadcast News Transcription Task (NAB, 60,000 words) [80-84].

2-dimensional HMM

- The speech spectrum is viewed as a Markov process (Weber et al,2000)

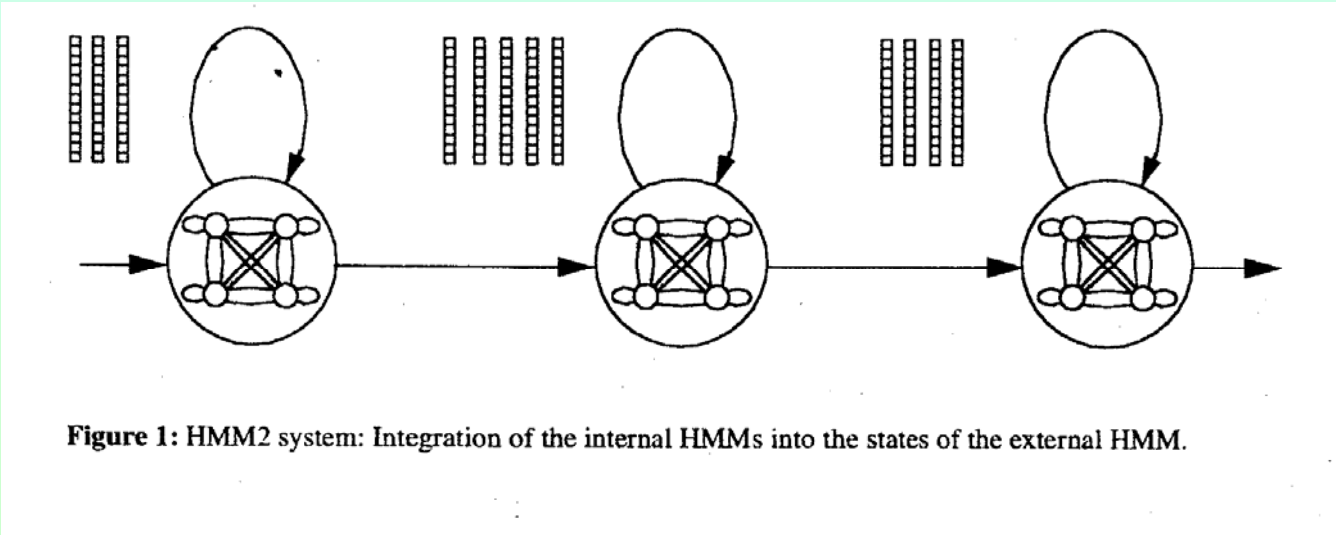
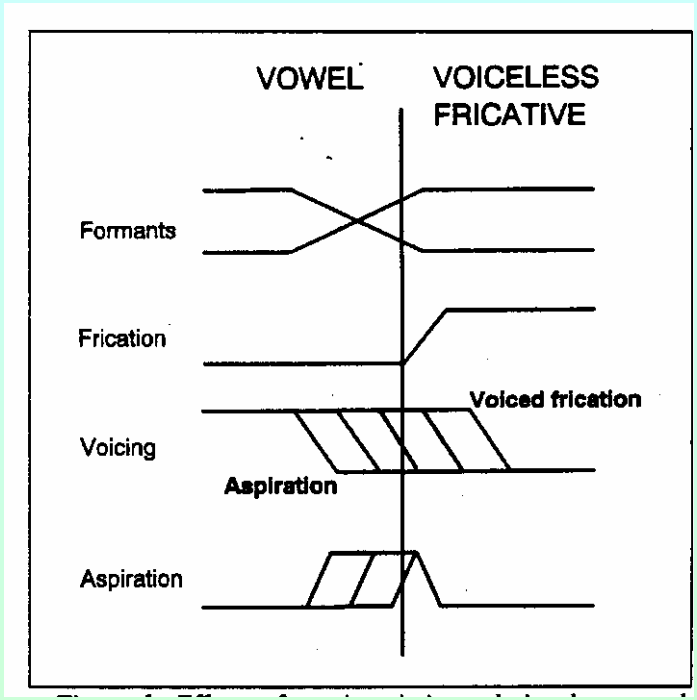


Figure 1: HMM2 system: Integration of the internal HMMs into the states of the external HMM.

Articulatory Inspired Modeling

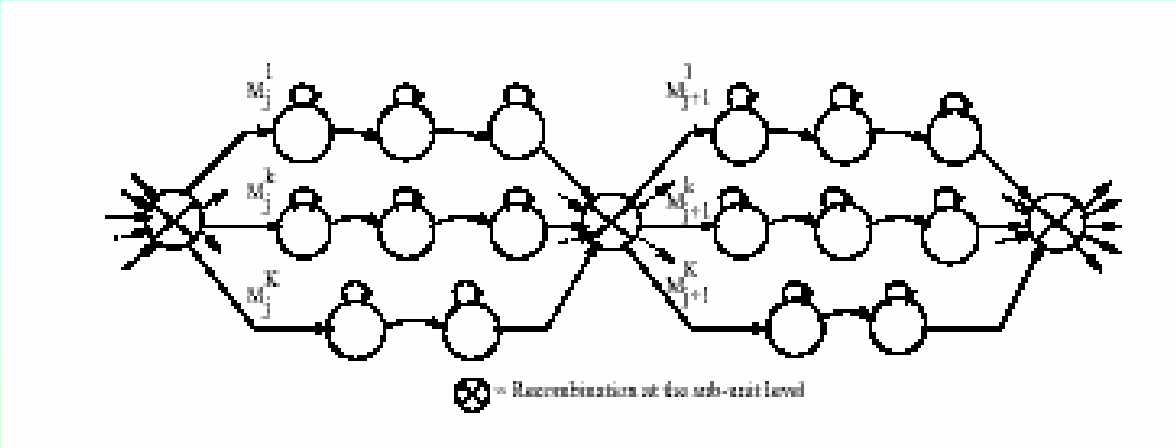
- Variation in articulator synchrony cause large acoustic variability
  - Ex. Transition region in boundary vowel - unvoiced fricative
    Which is first? Devoicing: Aspiration
    Closure: voiced fricative
- Linear trajectories in the articulatory domain are transformed to non-linearity in the spectral/cepstral domain
  - Should be easier to model coarticulation in the articulatory domain
  - Transformation to different physical size



Blomberg (1991)

Multi-stream Systems

- Dupont, Bourlard (1997)
- Separate decoding for feature subsets



Bayesian Networks

- Hidden Feature Modeling (Livescu et al, 2003)

Use of Prosody and Duration

- Carries semantic, stress, and non-linguistic information
  - Several information sources are superimposed
- Not fully synchronized to the articulation
  - Multi-stream technique would help
- Small improvement reported
  - 1% (Chen et al, 2003)

9.9 Case Study: Whisper

- Microsoft's general-purpose speaker-independent continuous speech recognition engine
  - MFCC + Delta + Acceleration
  - Cepstral Normalisation to eliminate channel distortion
  - Three-state phone models
  - Lexicon: mainly one pronunciation per word
  - Speaker adaptation using MAP and MLLR (phone-dependent classes)
  - Language model: Trigram (60 000 words) or context-free grammar
  - Performance: 7% WER on DARPA dictation test

Ch 10 Environmental Robustness

- The Acoustical Environment
- Acoustical Transducers
- Adaptive Echo Cancellation
- Multimicrophone speech enhancement
- Environment Compensation Preprocessing
- Environmental Model Adaptation
- Modeling Nonstationary Noise

10.1 The Acoustical Environment

- Additive Noise
- Reverberation
- A Model of the Environment

10.1.1 Additive Noise

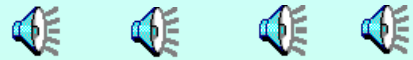
- Stationary - non-stationary
- White - colored
  - Pink noise
- Environment - speaker
- Real - simulated
  - The speaker may change his voice when speaking in noise (The Lombard effect)
  - Reported recognition experiments are mainly performed in simulated noise - do not capture this effect

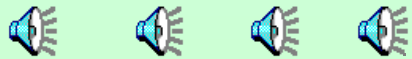
10.1.2 Reverberation

- Sound reflections from walls and objects in a room are added to the direct sound.
- Recognition systems are very sensitive to this effect
- Strong sounds mask succeeding weak sounds
- Reverberation radius - the distance from the sound source where the direct and the far sound fields are equal in amplitude
- Typical office
  - reverberation time up to 100 ms
  - reverberation radius 0.5 m

Environments

- Office - 200 speakers
  - at least 4 different rooms (close and far wall)
  - close talk, hands-free, medium distance (0.75 m), far distance (2 m)
- Public Place - 200 speakers
  - at least 2 locations: hall > 100m^2 and outdoors
- Entertainment - 75 speakers
  - at least 3 different living rooms with radio on/off,
- Car - 75 speakers
  - middle or upper class car
    - VW Golf, Opel Astra, Mercedes A Class
    - Ford Mondeo, Mercedes C Class, Audi A6
  - motor on/off, city 30-70, road 60-100, highway 90-130 km/h
- Children
  - 50 speakers
    - children's room

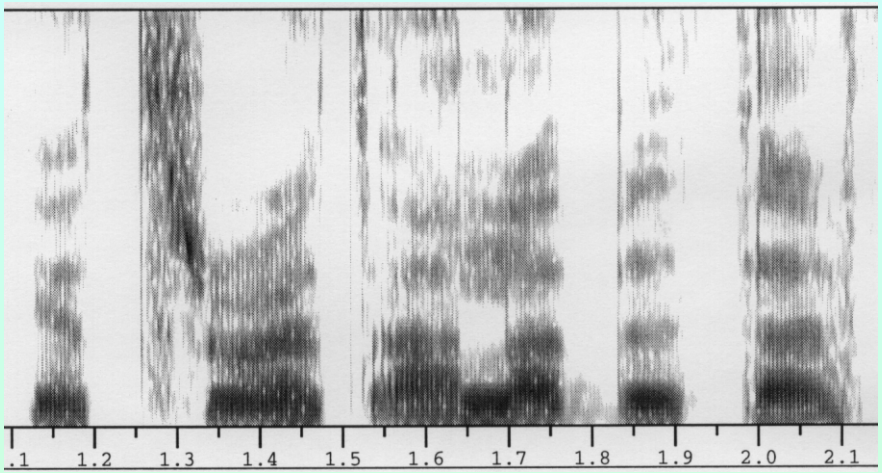




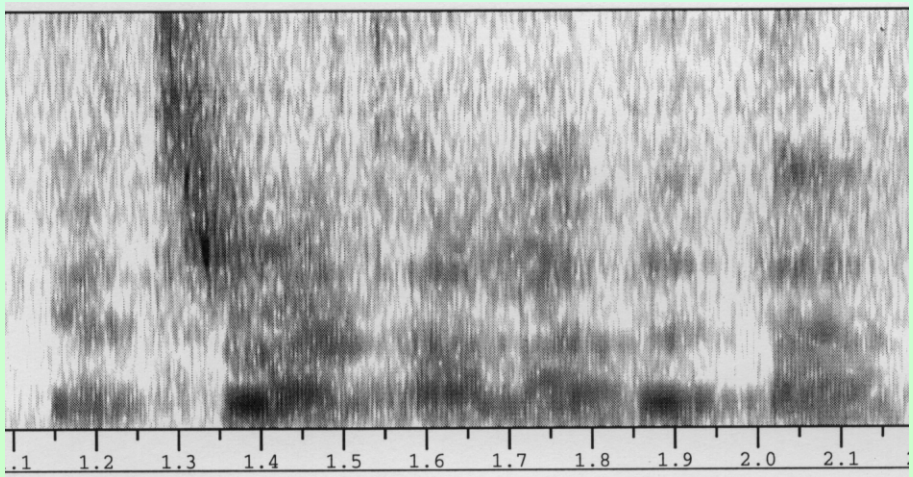
Near and far distance microphones

Stereo recording 2 microphones in quiet office

Headset

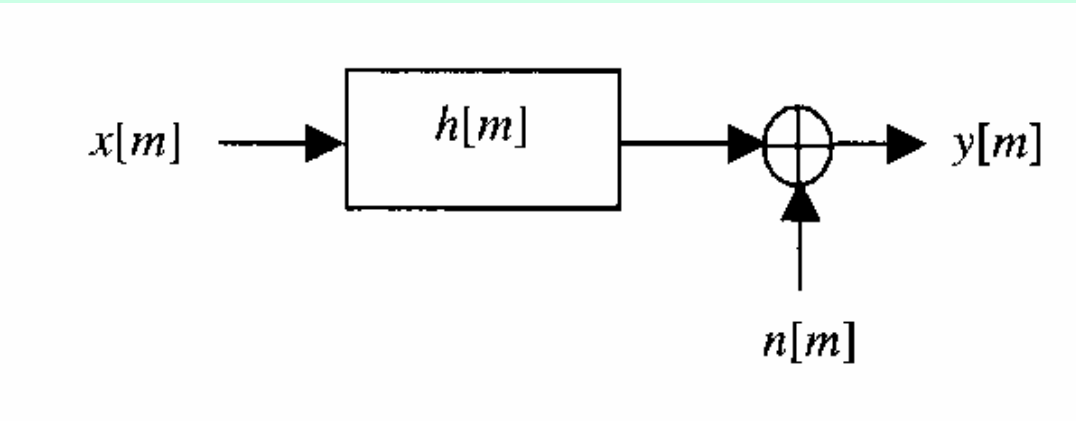


3 m distance



10.1.3 A Model of the Environment

- A model of combined noise and reverberation effects



Simulated Effect of Additive Noise

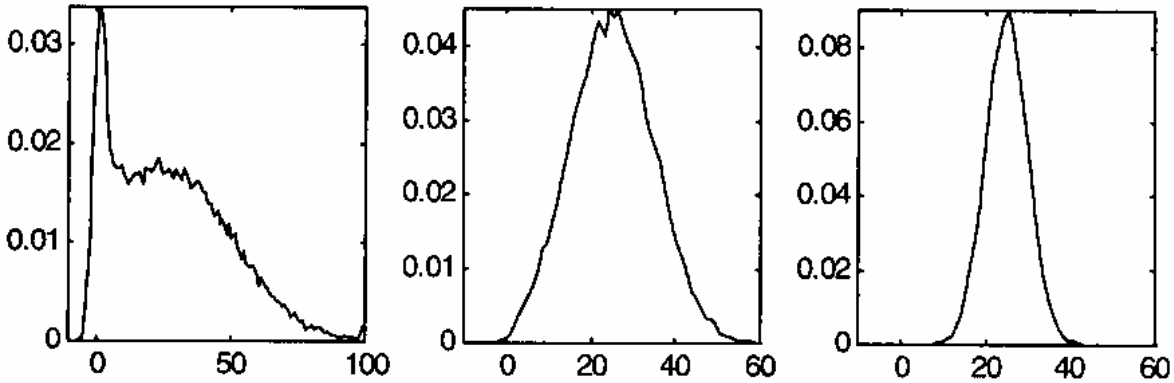


Figure 10.3 Distributions of the corrupted log-spectra y of Eq. (10.22) using simulated data. The distribution of the noise log-spectrum n is Gaussian with mean 0 dB and standard deviation of 2 dB. The distribution of the clean log-spectrum x is Gaussian with mean 25 dB and standard deviations of 25, 10, and 5 dB, respectively (the x-axis is expressed in dB). The first distribution is bimodal, whereas the other two are approximately Gaussian. Curves are plotted using Monte Carlo simulation.

10.2 Acoustical Transducers

- Close-talk and far field microphones
  - Close-talk
    - background noise is attenuated
    - sensitive to speaker non-speech sounds
    - positioning is critical
      - mouth corner recommended
      - plosive bursts may saturate the mic signal if right in front
  - Far field
    - picks up more background noise
    - positioning less critical
- Most popular type: condenser microphone
- Multimicrophones - Microphone Arrays
  - Adjustable directivity

10.3 Adaptive Echo Cancellation

- The LMS Algorithm
- Convergence Properties of the LMS Algorithm
- Normalized LMS Algorithm
- Transform-Domain LMS Algorithm
- The LRS Algorithm

10.4 Multimicrophone Speech Enhancement

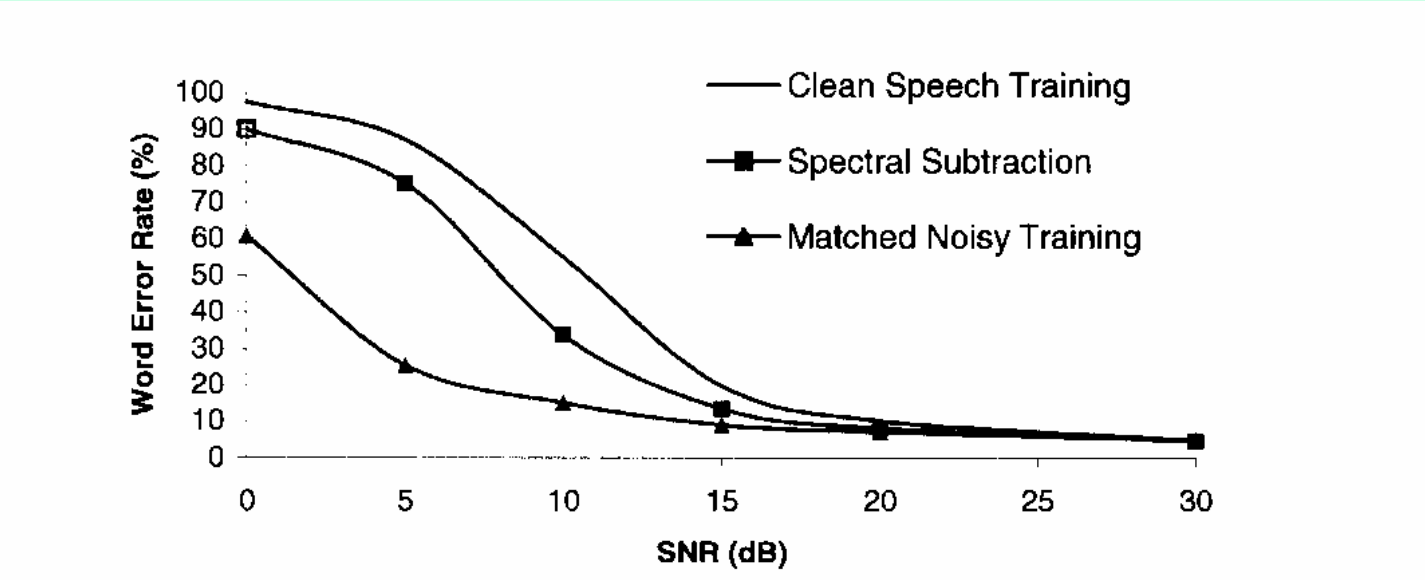
- Microphone Arrays
- Blind Source Separation

10.5 Environment Compensation Preprocessing

- Spectral Subtraction
- Frequency Domain from Stereo Data
- Wiener filtering
- Cepstral Mean Normalization (CMN)
- Real-Time Cepstral Normalization
- The Use of Gaussian Mixture Models

10.5.1 Spectral Subtraction

- The output power spectrum is a sum of the signal and the noise power spectra
- The noise spectrum can be estimated when there is no signal present and be subtracted from the output spectrum
- Musical noise in the generated speech signal at low SNR due to fluctuations



Noise Removal

- Frequency Domain MMSE from Stereo Data
  - Minimum mean square correction spectrum is estimated from simultaneously recorded noise-free and noisy speech
- Wiener Filtering
  - Find filter to remove the noisy signal
  - Needs knowledge of both noise and signal spectra
  - Chicken and egg problem

10.5.4 Cepstral Mean Normalization (CMN)

- Subtract the average cepstrum over the utterance from each frame
- Compensates for different frequency characteristics
- Problem
  - The average cepstrum contains both channel and phonetic information
  - The compensation will be different for different utterances
  - Especially for short utterances (< 2-4 sec)
- Still provides robustness against filtering operations
  - For telephone recordings, 30% relative error reduction
  - Some compensation also for differences in voice source spectra

10.5.5 Real-Time Cepstral Normalization

- CMN is not available before utterance is finished
  - Disables recognition output before end is reached
- Use a sliding cepstral mean over the previous frames for subtraction (time constant around 5 sec)
- Or use another filter, such as RASTA, which performs a bandpass filter (2- 10 Hz) on each filter amplitude envelope

10.5.6 The Use of Gaussian Mixture Models

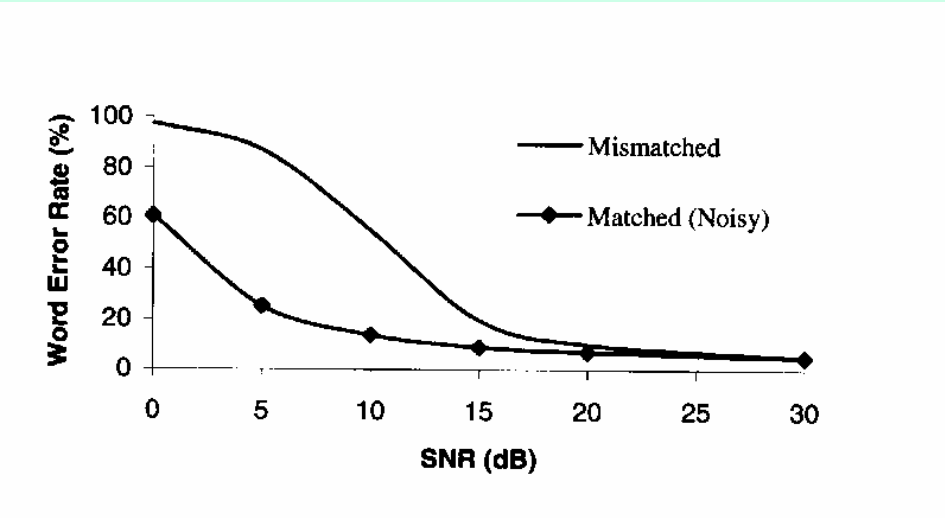
- Account for the fact that different frequencies are correlated
  - Avoids non-speech-like spectra
- Model the joint pdf of clean and noisy speech as a Gaussian mixture
- For each mixture component *k*, train the correction between clean and noisy speech using stereo recordings
- Pick the mixture that maximizes the joint probability of the clean and noisy speech cepstra
- Clean cepstrum estimate: $\hat{\mathbf{x}}_{ML} = \mathbf{C}_k \mathbf{y} + \mathbf{r}_k$
- No performance given

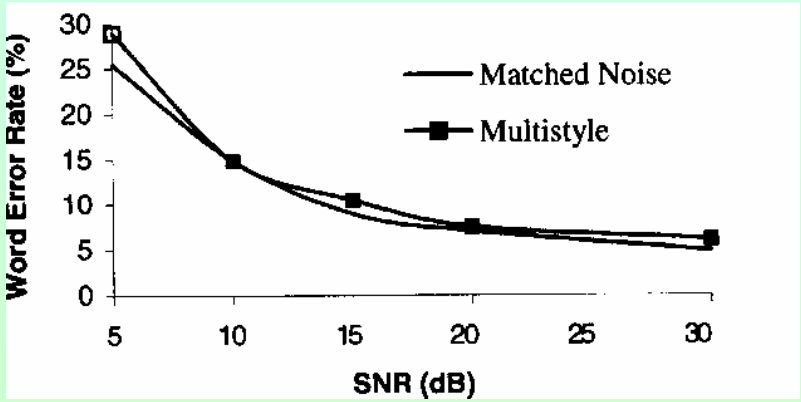
10.6 Environmental Model Adaptation

- Retraining on corrupted Speech
- Model Adaptation
- Parallel Model Combination
- Vector Taylor Series
- Retraining on Compensated Features

10.6.1 Retraining on Corrupted Speech

- If the distortion is known, then new models can be retrained on transformed non-distorted training data (noise added, filtering)
- Several distortions can be used in parallel (multistyle training)





10.6.2 Model Adaptation

- Same methods possible as for speaker adaptation (MAP and MLLR)
  - MAP requires large adaptation data - impractical
  - MLLR needs ca 1 min
- MLLR with one regression class and only bias works similarly to CMN but
  - Combined speech recognition and MLLR estimation of the distortion
  - Slightly better than CMN, especially for short utterances
  - Slower than CMN since two-stage procedure and model adaptation as part of recognition

10.6.3 Parallel Model Combination

- Noisy speech models = speech + noise models
  - Gaussian distribution converts into Non-Gaussian distribution (Cf Ch 10.1.3)
  - No problem, a Gaussian mixture can model this
  - Non-stationary noise can be modelled by having more than one state at the cost of multiplying the total number of states

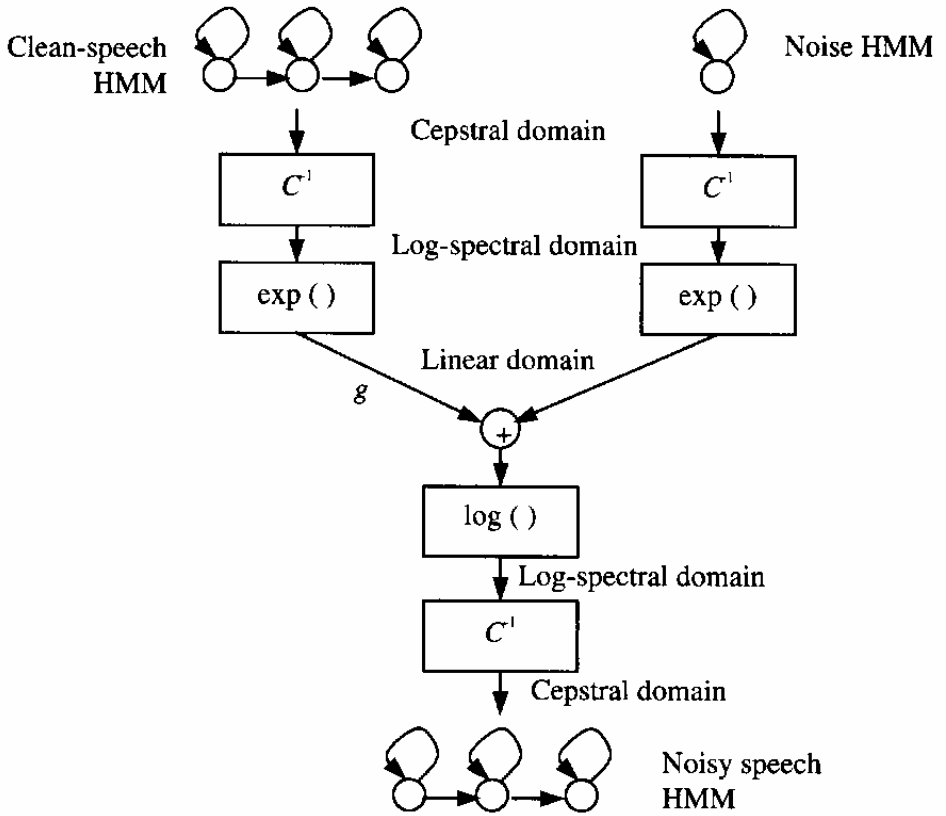


Figure 10.33 Parallel model combination for the case of one-state noise HMM.

10.6.4 Vector Taylor Series

- Use Taylor series expansion to approximate the nonlinear relation between clean and noisy speech
- New model means and covariances can be computed

10.6.5 Retraining on Compensated Features

- The algorithms for removing noise from noisy speech are not perfect
- Retraining can compensate for this

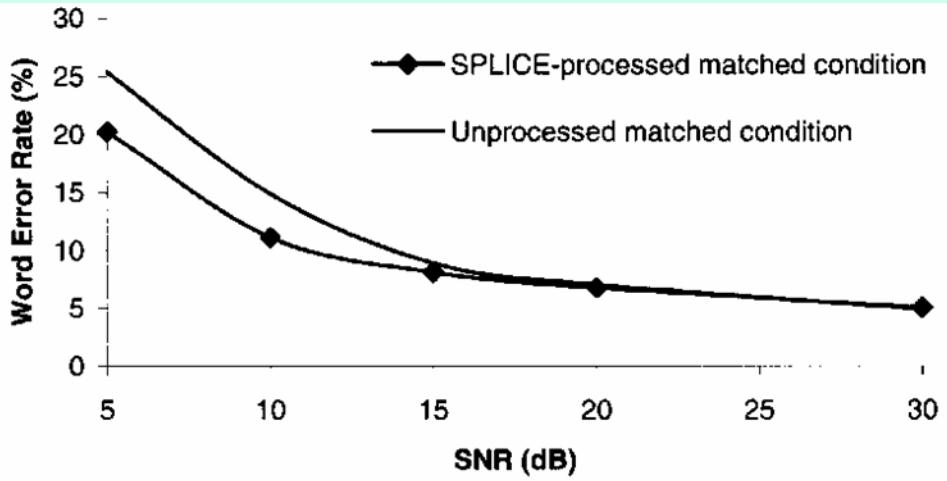


Figure 10.36 Word error rates of matched-noise training without feature preprocessing and with the SPLICE algorithm [21] as a function of the SNR in dB for additive white noise. Whisper is trained as in Figure 10.30. Error rate with the mixture Gaussian model is up to 30% lower than that of standard noisy matched conditions for low SNRs while it is about the same for high SNRs.

10.7 Modeling Nonstationary Noise

- Approach 1- Explicit noise modeling
  - Include non-speech labels in the training data
  - Perform training
  - Update the transcription using forced alignment where optional noise is allowed between words
  - Retrain
- Approach 2 - Speech/noise decomposition during recognition
  - 3-dimensional Viterbi
  - Computationally complex