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
- Adaptive Techniques – Minimizing Mismatches
- Confidence Measures: Measuring the Reliability
- Other Techniques than HMM
  - Alternatives / Complementary

- Performance illustrations on Microsoft’s Whisper system (≤ 2000)
  - Speaker-independent 60 000-word dictation, 300 speakers for acoustic model training, 2 billion words for training the language model
9.1 Variability in the Speech Signal

- **Context**
  - Linguistic
    - Homonyms: same pronunciation but spelling and meaning dependent on word context
    - "The two apples are too small. And expensive to buy."
  - Acoustic
    - Coarticulation, reduction effects
- **Speaking style**
  - Isolated words, read-aloud speech, conversational speech
- **Speaker**
  - Speaker-dependent, -independent, -adaptive
- **Environment**
  - Background noise, reverberation, transmission channel

9.2 How to Measure Speech Recognition Errors

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

Word error rate = \(\frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{No. of words in the correct sentence}}\)
Components of a Speech Recognition System

Components: 
- Acoustic Models
- Lexical Models
- Language Models

Signal Processing: 
16 kHz 100 Hz

Feature Extraction

Search & Match

Decoding

Recognized words

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

Effect of sampling rate on the performance

<table>
<thead>
<tr>
<th>Sampling rate</th>
<th>Relative Error-rate Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 kHz</td>
<td>Baseline</td>
</tr>
<tr>
<td>11 kHz</td>
<td>+10%</td>
</tr>
<tr>
<td>16 kHz</td>
<td>+10%</td>
</tr>
<tr>
<td>22 kHz</td>
<td>+0%</td>
</tr>
</tbody>
</table>

- Children’s speech recognition benefit from higher sampling rate

9.3.2 End-Point Detection

- Explicit detection
  - Two-class pattern classifier selects intervals to be recognised
    - In live speech recognition only
    - Not required in speech corpora – normally 1 utterance/file
  - 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, included non-speech intervals can be rescued by the recogniser
    - Dynamically adaptive algorithm (EM) better than fixed threshold
- Implicit detection by the recognizer
  - Better than explicit
  - But high computational demands
9.3.3 MFCC and Its Dynamic Features

- Mel Frequency Cepstral Coefficients
- The most popular feature type for speech recognition
- 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 $c_k$ often replaced by energy
  - 13th-order 40 ms 1st order deltas $\Delta c_k = c_{k+1} - c_{k-1}$
  - 13th-order 2nd order deltas $\Delta \Delta c_k = \Delta c_{k+1} - \Delta c_{k-1}$
- Deltas are often computed as regression lines

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Rel. Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>13th-order LPCC</td>
<td>Baseline</td>
</tr>
<tr>
<td>13th-order MFCC</td>
<td>+10%</td>
</tr>
<tr>
<td>16th-order MFCC</td>
<td>+0%</td>
</tr>
<tr>
<td>+ 1st and 2nd order deltas</td>
<td>+20%</td>
</tr>
<tr>
<td>+ 3rd order deltas</td>
<td>+0%</td>
</tr>
</tbody>
</table>

9.3.4 Feature Transformation

- The extracted acoustic features may be transformed in order to
  - Increase class separability
    - Principal Component Analysis (PCA)
    - Linear Discriminant Analysis (LDA)
  - Reduce influence of
    - noise, distortion (Ch 10)
    - speaker variability
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
    - Compare with PCA, whose eigenvectors represent directions with maximum class independent variability
PCA vs LDA
PCA finds directions with maximum class-independent variability
LDA finds directions with maximum class discrimination

LDA better for recognition

Support Vector Machines

- Expand the acoustic feature space to higher dimensionality in which the classes are linearly discriminable
- First use in speech technology: binary classifier in speaker verification
  - Extended to multiple classes in speech recognition

(Campbell et al., 2006)
9.3.4 Feature Transformation: VTLN

- VTLN: Frequency warping for Vocal Tract Length Normalisation
- Compression/expansion of the frequency axis to account for varying vocal tract size
- Two warping techniques
  - Modify center frequencies of filters in filter bank
  - Direct transformation in the MFCC domain (matrix multiplication)
- In theory, scaling should be phoneme dependent
  - Phoneme-independent scaling used in practice, works reasonably well.
- Warp factor range
  - 0.8 – 1.25 between adults
  - 1.0 – 1.7 children against adults
- Performance
  - Up to 10% relative error reduction among adult speakers
  - 50%-90% reduction for children’s speech using adult training speakers

9.4 Phonetic Modeling – Selecting Appropriate Units

- What is the best base unit for a continuous speech recogniser?
  - The base units: the inventory of trained models
  - A spoken utterance is composed of a sequence of base units
- 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 modeled 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 coarticulation is captured
  - Cross-word coarticulation is not
    - Requires word-pair training
  - Very large number, large vocabulary training unrealistic

- Syllable
  - Close tying with prosody (stress, rhythm)
  - Large number (30,000 in English), Cross-syllable coarticulation not captured

- Phone
  - Low number (around 50)
    - Very sensitive to coarticulation

- Context-dependent phone (triphone, diphone)
  - Captures coarticulation from adjacent phones
    - High number of triphones (125,000)

- Subphone
  - Captures fluctuation within the phone
    - Very high number – tying necessary
    - Can also be implemented as states in phone models

9.4.2 Context Dependency

- Triphones capture the dependence from immediately neighboring phonemes

- Dependence not captured:
  - Certain coarticulation
    - Phones at longer distance (e.g., lip rounded, retroflex, nasal)
    - Across word boundaries (unless cross-word triphones)
  - Stress information (mostly in duration and fundamental frequency)
    - Lexical stress (import vs. import)
    - 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, (~ States in Tied-state Triphones)
  - 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

- Here, the first two states can be tied:
  - The output distributions are similar

Predict unseen triphones

- Important since training data is always too small to reliably estimate infrequent triphones
- Which senones (states) should represent a triphone that does not exist in the training data?
- Select using decision tree
- CART-designed
Unit performance comparison

<table>
<thead>
<tr>
<th>Units</th>
<th>Rel. Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context-independent phone Baseline</td>
<td>Baseline</td>
</tr>
<tr>
<td>Context-dependent phone</td>
<td>+25%</td>
</tr>
<tr>
<td>Clustered triphone</td>
<td>+15%</td>
</tr>
<tr>
<td>Senone</td>
<td>+24%</td>
</tr>
</tbody>
</table>

Relative error reduction for different modeling 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 (LTS) conversion not good for English
  - Need trainable LTS converter
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 somany. The vowel *eai* is more likely to flap, thereby having a higher transition probability into *lai*.](image)

9.5.1 Choice of HMM Output Distributions

- Discrete, continuous, or semicontinuous HMM?
  - If training data is small, use DHMM or SCHMM
  - Gaussian mixture components in CHMM
    - Normally diagonal covariance matrix
    - The number of components is important, the optimum value depends on the training data size
    - With sufficient training data: CHMM (20 components) reduces SCHMM error by 15-20%
9.6 Adaptive Techniques – Minimizing Mismatches

- There is always mismatch between training and recognition conditions. Compensation techniques:
  - Normalization
    - Adjust the acoustic feature vectors of the unknown utterance to remove non-phonetic information
  - Adaptation
    - Minimize the mismatch dynamically with little calibration data
    - Supervised
      - knowledge of the correct identity
    - Unsupervised
      - Use the recognition result (which may contain errors)

9.6.1 Maximum a Posteriori (MAP)

- A new model is estimated using the training data interpolated with old information about the model
  \[
  \hat{\mu}_b = \frac{\tau_b \mu_{mx} + \sum_{i=1}^{T} \zeta(i,k) x_i}{\tau_b + \sum_{i=1}^{T} \zeta(i,k)}
  \]

- \(\tau_b\) is a balancing factor between the prior mean and the ML estimate. Can be a constant for all Gaussian components
- Similar for the covariance estimation
- Limitations
  - The prior model needs to be accurate
  - Needs observations for all models
9.6.2 Maximum Likelihood Linear Regression (MLLR)

- Models in the same regression class share the same transform
- Models not existing in the adaptation data are updated if models belonging to the same regression class are in the data
- Linear regression functions transform mean and covariance for maximizing the likelihood of the adaptation data
- Mean transformation:
  \[
  \overline{\mu}_k = A_c \mu_k + b_c
  \]
  - \(A_c\) is a regression matrix, \(b_c\) is an additive vector for regression class \(c\)
  - \(A\) and \(b\) can be estimated in a similar way as when training the continuous observation parameters
- Iteration for optimization
- If little training data, use few regression classes
- Can adapt both means and variances
- Does not adapt transition probabilities

MLLR adaptation illustration

- The transform for a class is optimized to maximize the likelihood of the adapted models to generate the adaptation data
CMLLR – Constrained MLLR

- The means and the covariances are modified by the same transformation matrix, common for all models
- Equivalent to multiplying the input feature vector by the inverse matrix
  - In which case compensation for the Jacobian determinant is required
  - Straightforward to implement

Speaker-Adaptive Training (SAT)

- Problem in speaker independent models:
  - Large model variances due to inter-speaker variability
- Solution: Speaker Adaptive Training
  - Removes the (large) inter-speaker variability from the models. Only the (much smaller) average intra-speaker variability is included.
  - Technique: MLLR adaptation “transforms” every speaker to an average position before training
  - The model variances are decreased, reducing errors 5-10% vs MLLR alone
  - Requires adaptation during recognition
MLLR performance

<table>
<thead>
<tr>
<th>Models</th>
<th>Relative Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHMM Baseline</td>
<td></td>
</tr>
<tr>
<td>MLLR on mean only</td>
<td>+12%</td>
</tr>
<tr>
<td>MLLR on mean and variance</td>
<td>+2%</td>
</tr>
<tr>
<td>MLLR SAT</td>
<td>+8%</td>
</tr>
</tbody>
</table>

One context-independent regression class for all context-dependent phones with same mid unit

9.6.3 MLLR and MAP Comparison

- MLLR better for small adaptation data, MAP is better when the adaptation data is large. Combined MLLR+MAP best in both cases
Eigenvoices

- Positions the new speaker in a speaker space
- The dimensions of this space are the mean elements of each training speaker’s models (supervector)
- The dimensionality is reduced by PCA
- Speaker adaptation consists of finding the test speaker’s position in this space
  - Very little adaptation data needed

9.6.4 Clustered Models

- A single speaker- and environment- independent model often has too much variability for high performance recognition and adaptation
- If the training data is sufficiently large, cluster it into smaller partitions
  - Gender, Microphone, Background noise, Speech style
- Gender-dependent models can reduce WER by 10%
- Refined speaker clustering can reduce it further, but not as much
- Environment clustering and adaptation in Chapter 10
Trends

Citation Frequency of Adaptation/Recognition Algorithms in Interspeech proceedings

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 \[ P(W|X) = \frac{P(W)P(X|W)}{P(X)} \sum \frac{P(W)P(X|W)}{P(W)} \]
- Good confidence estimator if the denominator is not ignored
- Represent the denominator \( P(X) \) by a general-purpose recognizer
  - E.g. phoneme recognizer
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
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

Mixture component sequences that never occurred in the same utterance during training, are allowed during recognition.
Standard HMMs allow every frame in an utterance to come from a different speaker.

Example:
- Northern accent
- Southern accent

2-dimensional HMM

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

Figure 1: HMM3 system: Integration of the internal HMMs into the states of the external HMM.
Articulatory Inspired Modeling

- Variation in articulator synchrony cause large acoustic variability
- If more than one articulator position is changed in a phoneme boundary, variability in their relative timing will produce different intermediate segments
  - Ex. Transition region in boundary vowel - unvoiced fricative
  - Devoicing before closure: Aspiration
  - Closure before devoicing: 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

Multi-stream systems

- Dupont, Bourlard (1997)
- Separate decoding for feature subsets
- Accounts for timing asynchrony within the feature vector
  - E.g. between different frequency regions
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
Research progress history

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, 1,000 words), the Air Travel Information System spontaneous speech understanding task (ATIS, 2,000 words), the Wall Street Journal dictation task (WSJ, 20,000 words), and the Broadcast News Transcription Task (NAB, 60,000 words) [80-84].

How large training data to reach human listening performance?

*Extrapolated word error rates for increasing quantities of training data (Moore, Eurospeech 2003)*

Heard during a life-time

Saturation effect

Human performance
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
- Speaker
  - lip-smack, aspiration, multi-speaker environment (babble)
- 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
  - The microphone distance should be much smaller
- Typical office
  - reverberation time up to 100 ms
  - reverberation radius 0.5 m
Near and far distance microphones

Stereo recording 2 microphones in quiet office

Headset

3 m distance

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.

10.5.4 Cepstral Mean Normalization (CMN, CMS)

- 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). E.g. 1-phoneme utterance.
- Still provides robustness against filtering operations:
  - For telephone recordings, 30% relative error reduction.
  - Some compensation also for differences in voice source spectra.
RASTA – RelAtive SpecTrAl

- Hearing-inspired bandpass filtering of filterbank amplitude envelopes
- Removes long-term bias in the signal but leaves syllable rate modulation mainly unchanged

(Hermansky & Morgan, 1994)

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 models can be trained by distorting the training data in this way (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
  - 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 modeled by having more than one state at the cost of multiplying the total number of states

![Diagram of Parallel Model Combination]

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