Acoustic Modeling – Part 2
Outline

• Discrete versus Continuous HMMs
  – Parameter Tying
  – Pronunciation Variants
  – Speech Units
  – Context Dependent Acoustic Modeling
    • Bottom-Up vs. Top-Down Clustering
  – Clustering with Decision Trees

• Current Issues in Acoustic Modeling
  – Practical Issues with HMMs
    • Distances Between Model Clusters
  – Clustering of Context
    • Problems with Vocabulary Dependence
There are two different approaches to clustering:

- **Bottom-up clustering** (agglomerative): look for good combination of two classes into one

- **Top-down clustering** (divisive): look for good separation of a class into two subclasses

Both approaches result in a clustering tree:
1. Start with classes $C_i = \{ \text{Phone}_i \}$

2. Compare all class pairs: $C_i$ with $C_j$ ($j > i$)

3. If we find that $C_i$ and $C_j$ are "similar enough"
   – replace $C_i$ with $C_i + C_j$
   – remove $C_j$

4. Continue until satisfied.
## Clustering of Contexts (2)

### Second Idea for Context Tying:

**Unsupervised Clustering (top down):**

1. Start with class $C_0 = \{ context_1, context_2, ..., context_n \}$
2. Anticipate all possible splits of every class $C_i$ into two subclasses
3. If we find that it is good idea to split $C_i$ then
   - replace $C_i$ with its two subclasses
4. Continue with step 2 until satisfied

### Big Problem:

If we start with $n$ different contexts of the same phoneme then there are $2^n$ possible separations!

- Most real-world cases have hundreds of contexts
- This makes the algorithm not applicable
What means a „good“ split / merge?

Use distance measure between model clusters to ...

• Decide if a class separation or combination is "good"

• Find out which separation/combination is best
Distances between Model Clusters

Continuous parametric models:

\[ d(f,g) = \int \min(f(x),g(x)) \]

or Kullback-Leibler distances,

and others: \[ KL(f,g) = \sum f(x_i) \log f(x_i) / g(x_i) \]

In general (but typically for nonparametric models):

**Entropy-distance:** \[ d(f,g) = H(f+g) - 1/2 \cdot H(f) - 1/2 \cdot H(g) , \]
where \( H(f) \) is the entropy of the function \( f \), and \( f+g \) is the combined model.

Combining two models \( \rightarrow \) lose parameters \( \rightarrow \) lose information

The entropy distance measures the loss of information.

**Goal:** Lose as little information as possible! \( \rightarrow \) Minimize \( d \)
Discrete Entropy Distance

Remember:
Semi-continuous and discrete HMMs are represented by discrete distributions

For a discrete distribution \( f[i] \) the entropy:
\[
H(f) = - \sum_i f[i] \log_2 f[i] = \sum_i f[i] \log_2 \left( 1/f[i] \right)
\]
\[
d(f,g) = H(f+g) - 1/2 \cdot H(f) - 1/2 \cdot H(g)
\]

Obvious:
If \( f=g \) then \( H(f) = H(g) = H(f+g) \), thus \( d(f,g) = 0.0 \)
If \( f = \{ 1 \ 0 \}, \ g = \{ 0 \ 1 \} \) then \( H(f) = H(g) = 0, H(f+g) = 1, d(f,g) = 1.0 \)

Example: \( f = \{ 1/2 \ 1/2 \}, \ g = \{ 3/4 \ 1/4 \}, \ f+g = \{ 5/8 \ 3/8 \} \)

\[
H(f) = \frac{1}{2} \cdot \log_2(2) + \frac{1}{2} \cdot \log_2(2) = 1.0
\]

\[
H(g) = \frac{3}{4} \cdot \log_2(4/3) + \frac{1}{4} \cdot \log_2(4/1) = 0.811
\]

\[
H(f+g) = \frac{5}{8} \cdot \log_2(8/5) + \frac{3}{8} \cdot \log_2(8/3) = 0.954
\]
\[
\Rightarrow d(f,g) = 0.049
\]
Weighted Discrete Entropy Distance

Problem:
Speech examples are not equally distributed among models (some (poly-)phones are more frequent than others)

\[ M_I : \text{Model trained with many examples (=robust)} \]
\[ M_f : \text{Model trained with few examples (=unreliable)} \]
\[ M_{f+} : \text{Model trained with few but more examples than } M_f \]

Combining \( M_I \) with \( M_f \) should have a minor impact on the distance than combining \( M_I \) with model \( M_{f+} \)

Solution:
Weight the model entropy by number of training samples, so the commonly used entropy distance is:

\[ d(f,g) = (n_f+n_g) \cdot H(f+g) - n_f \cdot H(f) - n_g \cdot H(g) \]
Context Clustering after Kai-Fu Lee

1. Train semicontinuous models for all three states of each triphone, e.g. triphone = T(AE,K)-b T(AE,K)-m T(AE,K)-e

2. Initialize a context class for every triphone (a class is defined by three distributions: e.g. T_{17}-b T_{17}-m T_{17}-e)

3. Compute all distances between different context classes of same phone:
   \[ d(C_i,C_j) = E(C_i-b,C_j-b) + E(C_i-m,C_j-m) + E(C_i-e,C_j-e), \]
   where \( E \) is the weighted entropy distance

4. Replace the two classes with the smallest distance by their combination

5. Try to improve distance by moving any element from any class into any other class

Continue with step 3 while end criterion is not met

Note: This algorithm is completely data driven. Step 5 is expensive but important.
Generalized Triphones vs. Senones

Kai-Fu Lee's algorithm produces generalized triphones:

A better approach (M. Hwang) produces generalized subtriphones (senones):
Problems with Vocabulary Dependencies

Example Scenario:
During the training we have seen the phoneme \( P_1 \) in the contexts

\[
P_1(P_2,P_3), \quad P_1(P_4,P_5), \quad P_1(P_6,P_7), \quad P_1(P_8,P_9).
\]

After clustering we have found the classes:

\[
C_1 = \{ P_1(P_2,P_3), P_1(P_4,P_5) \} \quad \text{and} \quad C_2 = \{ P_1(P_6,P_7), P_1(P_8,P_9) \}
\]

During testing we would like to recognize the word with phoneme sequence:

\[
P_3P_1P_7
\]

Problem: Do we use \( C_1 \) or \( C_2 \) to model \( P_1(P_3,P_7) \)?
Clustering with Decision Trees

Approaches to achieve vocabulary independency:
1) If the test vocabulary contains an untrained context $m(l,r)$ then use the context independent model $m$ that was trained on all contexts
2) Use the model of a context class that is somehow "similar" to the unseen context

In general:
• If a context has not been seen during training then use some class further up in the hierarchy that was trained.
• To make a system independent of the vocabulary, we have to be able to find out in which context class it would have been clustered
• This approach discourages the usage of recorded data (there's not much during the test, and we don't know where it is)

Solution:
Build a decision tree that asks phonetic questions about the context
Clustering with Decision Trees

Example Decision Tree:

```
-1 = vowel?
  no
  A(P,M)→b
  A(L,F)→b
  A(L,U)→b

+1 = fricative?
  no
  A(P,M)→b
  A(L,U)→b

  yes
  A(O,N)→b
  A(I,T)→b
```

Clustering Algorithm:

1. Initialize one cluster containing all contexts
2. For all clusters / questions: compute distance of subclusters
3. Perform the split that get the largest distance (information gain)
4. Continue with step 2 until satisfied (number of clusters)
Typical optimal entropy distance during clustering:

How many clusters (models) do we want?

• Standard answer to many question about the amount of parameters: "As many as our CPU / memory can stand."
• Some kind of intelligent guess (based on experience)
• Number of samples per cluster does not fall below a certain threshold
• Use cross-validation set:
  - Separate training data into 2 (or more) subsets A,B, train models from A
  - When computing the distance between clusters $C_1$ and $C_2$:
    Compute the likelihood $P_1$ of all data from B that belong to $C_1$,
    compute the likelihood $P_2$ of all data from B that belong to $C_2$,
    and $P_{1+2}$ of all data from B that belong to $C_1$ or $C_2$ using the combined class $C_1+C_2$, define the distance as $(P_1 \cdot P_2) / P_{1+2}$
  - The likelihood gain of the split will not always be positive
Growing the Decision Tree

Acoustic Modelling I + II - 16
During training five contexts have been seen. These were clustered into three clusters. If we need to model the context $A(G,S)$ we will use:

- Left context of $A$ is a vowel? (-1 = vowel?) NO $\rightarrow$ G is not a vowel
- Right context of $A$ is a fricative (+1 = fricative?) YES $\rightarrow$ S is a fricative

$\Rightarrow$ use model A-b(4).

Problem: Where do the questions come from?
Where do the Questions come from?

- **Knowledge-based**
  - Expert defines “natural classes” based on *IPA classification*
  - Example list: Table 9.3 of [XH]
    - nasal: m n ng
    - velar: k g ng
    - labial: w m b p v
    - and 39 other classes

- **Automatically learned classes (e.g. Rita Singh, CMU)**
  - Provide phone names and feature properties
  - Use acoustic distance to cluster features
  - These become questions for context clustering

- **Random Selection of questions (IBM)**
  - Proved that the selection of questions is not critical
Question Sets for Decision Trees

- **Problem:** How to find good questions for decision tree
- **Answer:** Does the question set really matter?

- **Study 1:** Impact of different question sets
  - *IBM study:* As long as the question set allows variable enough separation, no significant differences

- **Question 2:** Impact of different trees based on same question set
  - *Siohan et al, IBM:* Randomly choose among the top-N best splits instead of always selecting the best split
  - Construct ensemble of ASR systems based on different trees
  - Single ASR systems do not significantly differ terms of WER but combined ASR results using ROVER gives big gain
  - Systems make different errors!
Rapid Portability: Acoustic Models

Input: Speech

Phone set & Speech data

Output: Speech & Text

Hello

Input: Speech

AM
Lex
LM
NLP / MT

TTS

สวัสดี ครับ
Output: Speech & Text
Rapid Portability: Data

Step 1:
- Uniform multilingual database (*GlobalPhone*)
- Build monolingual acoustic models in many languages
Step 2:
• Combine monolingual acoustic models to a set of multilingual (ML) “language independent” acoustic model
Rapid Portability: Acoustic Models

Step 3:
- Define mapping between ML set and new language
- Bootstrap acoustic model of unseen language

Input: Speech

Output: Speech & Text

NLP / MT

TTS

Hello

สวัสดี ครับ
Universal Sound Inventory

Speech production is independent from language  \( \Rightarrow \) IPA

1) IPA-based Universal Sound Inventory

2) Each sound class is trained by data sharing
   - Reduction from 485 to 162 sound classes
   - \( m,n,s,l \) appear in all 12 languages
   - \( p,b,t,d,k,g,f \) and \( i,u,e,a,o \) in almost all

![Diagram of vowel pronunciation](image)
**Problem:**
Context of sounds are language specific. How to train context dependent models for new languages?

**Solution:**
1) Multilingual decision context trees
2) Specialize decision tree by adaptation
Context Decision Trees

- Context dependent phones ($\pm n = \text{polyphone}$)
- Trainability vs. granularity
- Divisive clustering based on linguistic motivated questions
- One model is assigned to each context cluster
- Multilingual case: Should we ignore language information?
  - Depends on application
  - Yes in case of adaptation to new target languages
Polyphone Coverage

![Graph showing polyphone coverage across different languages](image)

- Weighted coverage of Portuguese monophones
- Weighted coverage of Portuguese triphones
- Weighted coverage of Portuguese quinphones

Multilingual Speech Processing, Schultz & Kirchhoff (ed.), Chapter 4, p.101
Rapid Language Adaptation

Model mapping to the target language
1) Map the multilingual phonemes to Portuguese ones based on the IPA scheme
2) Copy the corresponding acoustic models in order to initialize Portuguese models

Problem:
Contexts are highly language specific.
How to apply context dependent models to a new target language?

Solution:
1) Train a multilingual polyphone decision tree
2) Specialize this tree to target language using limited data
   (Polyphone decision tree specialization - PDTS)
Polyphone Decision Tree Specialization (1)

English Polyphone Tree
Polyphone Decision Tree Specialization (2)
Polyphone Decision Tree Specialization (3)

Multilingual Polyphone Tree
Polyphone Decision Tree Specialization (4)

Polyphones found in Portuguese
1. Tree Pruning:
Select from all polyphones only the ones which are relevant for the particular language.
2. Tree Regrowing

Further specialize the tree according to the adaptation data
Rapid Portability: Acoustic Model

Word Error rate [%]

Ø Tree  ML-Tree  Po-Tree  PDTS

0  00:15  00:15  00:25  00:25  00:25  01:30  16:30

69.1  57.1  49.9  40.6  32.8  28.9  19.6  19
Traverse and Analyze the Decision Tree

![Graph showing the entropy gain over the number of splits for different languages. The x-axis represents the number of splits, and the y-axis represents the entropy gain. The labels for the curves include:
- Summe aller Fragen
- Phonetische Kontextfragen
- Sprachenfrage Koreanisch
- Sprachenfrage Tuerkisch
- Sprachenfrage Kroatisch
- Sprachenfrage Japanisch
- Sprachenfrage Spanisch]
Current Research Issues in Acoustic Modeling

- **Data Collection, Lack of transcripts**
  
  "There's no data like more data."
  
  Example: Training with 20h of speech $\rightarrow$ 13% WER,
  
  with 80h of speech $\rightarrow$ 9% WER
  
  Today 5,000 hours audio material
  
  (with partly semi-automatically generated transcripts)

- **Signal preprocessing**
  
  (remove the unimportant, enhance the important)

- **Training techniques** (ML, MAP, discriminative training, ...)

- **Parameter tying** (what are acoustic atoms)

- **Usage of memory and CPU resources**

- **Robustness** (reduce effect of disturbances)

- **Adaptation** (keep learning and improving while in use)

- **Multilinguality** (recognizers in many languages)
Current Research Issues in AM: Multilinguality

Language independent recognizers
(in analogy to speaker independent)

Benefits:

• More training data for only little more parameters
• Same acoustic model can be trained with different languages
  ➔ more robust?
• Language Identification included, no other module necessary
• Rapid deployment of acoustic models to new target language
• Allows code-switching (= language switch within a sentence)

Problems:

• What is a good common set of phonemes (speech units)?
• How to decide which speech units are similar across languages?
• How to fight the "smearing" effect?
  (different appearances of same model)
Polyphones Types over Context Width for 9 Languages

- Chinese
- German
- French
- Croatian
- Spanish
- Turkish
- Japanese
- English
- Korean
- Chinese, Japanese, Korean (CH, JA, KO)
- German, French, Croatian, Spanish, Turkish (DE, FR, ES, TR)
- Croatian, Spanish, Turkish (HR, ES, TR)
- Chinese, Japanese, Korean, German, French, Croatian, Spanish, Turkish (CH, JA, KO, DE, FR, ES, TR)

Number of Sub-polyphone Types vs. Context Width
Number of Polyphones

- Depends on the language
- Length of context (triphones = ±1, quinphones = ±2, …)
- Number of mono-phone types (may vary between [30…150])
- Phonotactics (consonantal clusters, mora, …)
- Morphology
- Word segmentation
Universal Sound Inventory

Speech production is independent from language → IPA

1) IPA-based Universal Sound Inventory

2) Each sound class is trained by data sharing
   - Reduction from 485 to 162 sound classes
   - \( m, n, s, l \) appear in all 12 languages
   - \( p, b, t, d, k, g, f \) and \( i, u, e, a, o \) in almost all
Acoustic Model Combination

Word Error Rate [%]

- Croatian
  - Mono: 27
  - ML-Tag7500: 30
  - ML-Tag3000: 32
  - ML-Mix3000: 35

- Japanese
  - Mono: 13
  - ML-Tag7500: 14
  - ML-Tag3000: 15
  - ML-Mix3000: 20

- Spanish
  - Mono: 28
  - ML-Tag7500: 30
  - ML-Tag3000: 32
  - ML-Mix3000: 37

- Turkish
  - Mono: 20
  - ML-Tag7500: 21
  - ML-Tag3000: 21
  - ML-Mix3000: 29
Lack of Transcripts

- Projects such as EARS and GALE initiated the collection of vast amount of audio data (up to 5,000 hours by now)
- Will likely to become more rapidly (e.g. Jim Baker, 1 Mio hour plan)
- **Problem:** Can not be transcribed by human beings any more
- **Solution 1:** *Quick Transcription*
  - Use pre-existing recognizer to decode audio recordings
  - Ask humans to cross-check the output (in about 6 times real-time)
    - If hypothesis is correct → PASS
    - If hypothesis is close enough → CORRECT
    - If hypothesis is off → THROW AWAY
  - Still too expensive, also lose all bad hypos
- **Solution 2:** *Slightly Supervised Training*
  - Some kind of references are given (similar to close captions)
    - Step 1: Create a biased language model on these close captions
    - Step 2: Decode all audio recordings using this biased language model
      - Find speech portions of high “confidence” and train on those
  - Leads to significant improvements (e.g. GALE 1500hrs ~ 5-7% relative)
- **Solution 3:** *Unsupervised Training* – see lecture by Thang/Tim
Current Research Issues in AM: Signal Preprocessing

Minor Issues:
• What features to use?
  (Spectrum, Cepstrum, LPC, bottleneck features ...)

Major Issues:
• Normalization techniques for speaker, e.g. vocal tract lengths (VTLN), Speaker Adaptive Training, Articulatory Features
• Preprocessing "afterburner" (RASTA, LDA, HLDA, ...)
• Dynamic Features (higher order HMMs, formant shapes, ...)
• Decomposition (multidimensional HMMs, ...)
• Noise Reduction (echo canceling, car-noise reduction, ...)
• Speaker Segmentation and Clustering
Current Research Issues in AM: Robustness

In general: **Robustness** is stability against variations.

Variations that affect the recognition accuracy are:
- Speech itself (styles, speeds, dialects, spontaneity, ...)
- Background noise (car, cocktail party, street noise, music, ...)
- Channel effects (microphones, telephone, room specs, ...)

Current Efforts:
- Enhance the part that humans use to recognize speech
- Suppress those parts that are irrelevant (e.g. noise subtraction)
- Normalization (map different appearances of same thing to one appearance, e.g. VTLN)
Current Research Issues in AM: Adaptation

In general: Adaptation is modifying the parameters such that they fit better onto the current signal (= model adaptation) or modifying the signal such that it fits better onto the system's parameters (= feature space adaptation)

Most common adaptations:
• Adapt to the speaker (move speaker independent recognizer towards a speaker dependent)
• Adapt to the environment (make recognizer a bit more environment-dependent)

Reasons for using adaptation:
• Speaker or environment dependent recognizers are more precise
• Data sparseness: not enough data available to train a speaker dependent recognizer, adaptation can work with fewer data
Summary Acoustic Modeling (Part 1+2)

- Pronunciation Variants
- Context Dependent Acoustic Modeling
- From Sentence to Context Dependent HMM
- Speech Units
- Crossword Context Modeling, Problems
- Tying of Contexts
- Clustering of Context
- Bottom-Up vs. Top-Down Clustering
- Distances Between Model Clusters
- Problems with Vocabulary Dependencies
- Clustering with Decision Trees
- Some open questions in AM

Upcoming: Adaptation, Special problems
Thanks for your interest!