Summary

- We aim to capture the pronunciation variations of speakers by modeling the linkage between the states of articulation and the actual phones produced by a speaker.
- We propose using a mixture of discrete density functions for articulatory feature-based conditional pronunciation modeling (AFCPM), with one model representing vowels and another representing consonants. Verification scores are the weighted sum of the two models' outputs.
- Four articulatory properties (front-back, lip-rounding, place of articulation, and manner of articulation) were used for pronunciation modeling.
- Results show that dividing the articulatory properties into two groups is an effective means of solving the data-sparseness problem encountered in the training phase of AFCPM systems.

Articulatory Features (AFs)

- Articulatory features (AFs) are abstract classes that describe the movements and positions of different articulators during speech production.
- Four AFs were adopted for CPM:

Articulatory Properties	*Classes (AFs)*	*Number of Classes*
Front-back ($\mathcal{FB}$)	Silence, Front, Back, Nil	4
Rounding ($\mathcal{R}$)	Silence, Rounded, Not Rounded, Nil	4
Manner ($\mathcal{M}$)	Silence, Vowel, Stop, Fricative, Nasal, Approximant-Lateral	6
Place ($\mathcal{P}$)	Silence, High, Middle, Low, Labial, Dental, Coronal, Palatal, Velar, Glottal	10

Articulatory Feature Extraction

- At frame position t, 9 consecutive frames of MFCCs (X_t) centered at frame t were input to an MLP.

Manner labels

$$l_t^M = \arg\max_{m \in M} P(\text{Manner} = m \mid X_t)$$

Place labels

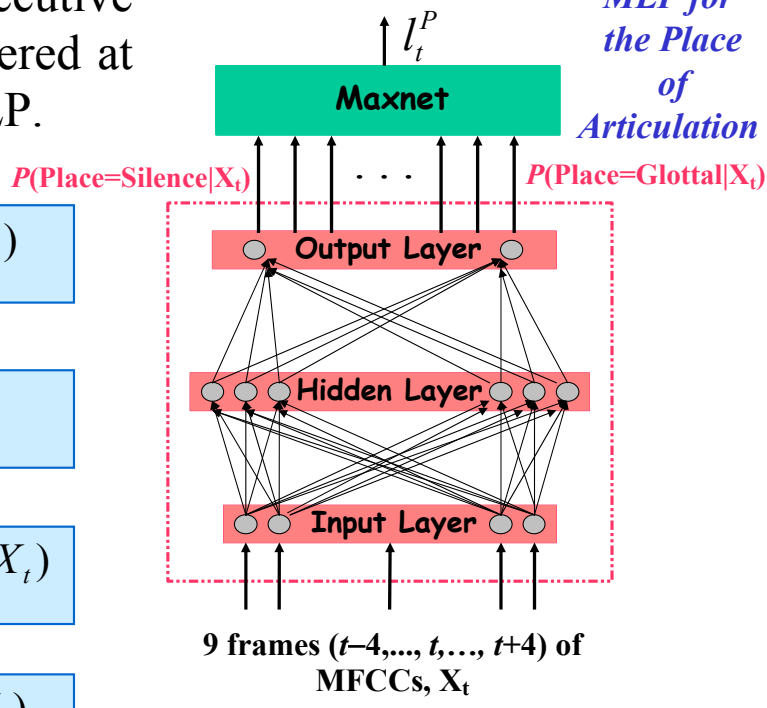
$$l_t^P = \arg\max_{p \in P} P(\text{Place} = P \mid X_t)$$

Front-back labels

$$l_t^{FB} = \arg\max_{fb \in FB} P(\text{FrontBack} = fb \mid X_t)$$

Lip-rounding labels

$$l_t^R = \arg\max_{r \in R} P(\text{LipRound} = r \mid X_t)$$



AF-Based CPM: Training

- To train the probabilistic models for speaker s, we compute the following for each of the phonemes:

$$P_s(m, p \mid q) = P_s(\text{Manner} = m, \text{Place} = p \mid \text{Phoneme} = q) = \frac{\text{No. of } \{m, p, q\} \text{ from the data of speaker } s}{\text{No. of } \{q\} \text{ from the data of speaker } s}$$

$$P_s(fb, r, p \mid q) = P_s(\text{FrontBack} = fb, \text{Rounding} = r, \text{Place} = p \mid \text{Phoneme} = q) = \frac{\text{No. of } \{fb, r, p, q\} \text{ from the data of speaker } s}{\text{No. of } \{q\} \text{ from the data of speaker } s}$$

AFCPM Training: Example

Given an 11-frame training utterance of speaker s, the speaker model corresponding to phoneme /t/ can be obtained as follows:

P(Manner=***Vowel***, Place=***Low*** | **/t/**) = 1/6,
P(Manner=***Silence***, Place=***Silence*** | **/t/**) = 4/6,
P(Manner=***Stop***, Place=***Coronal*** | **/t/**) = 1/6,
and all other entries are equal to 0.

Manner \ Place	Silence	High	Middle	Low	Labial	Dental	Coronal	Palatal	Velar	Glottal
Silence	4/6	0	0	0	0	0	0	0	0	0
Vowel	0	0	0	1/6	0	0	0	0	0	0
Stop	0	0	0	0	0	0	1/6	0	0	0
Fricative	0	0	0	0	0	0	0	0	0	0
Nasal	0	0	0	0	0	0	0	0	0	0
Approximant / Lateral	0	0	0	0	0	0	0	0	0	0

Frame, t	Phoneme, q_t	l^{m_t}	l^{p_t}
1	/t/	Vowel	Low
2	/t/	Silence	Silence
3	/t/	Silence	Silence
4	/t/	Silence	Silence
5	/t/	Silence	Silence
6	/t/	Stop	Coronal
7	/aa/	Vowel	Low
8	/aa/	Vowel	Low
9	/aa/	Vowel	Low
10	/aa/	Vowel	Low
11	/aa/	Vowel	Low

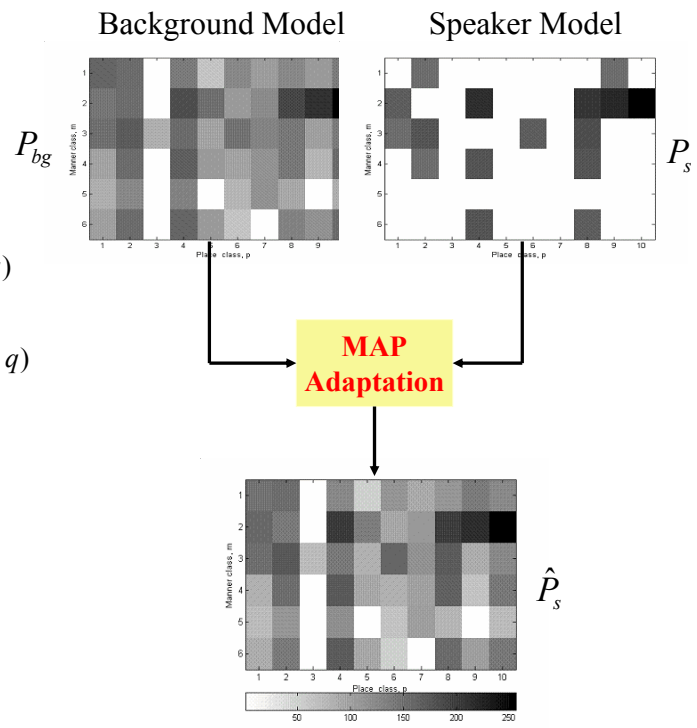
AF-Based CPM: MAP adaptation

- To overcome the data sparseness problem and enhance the coupling between the speaker models and background models, speaker models are adapted from background models:

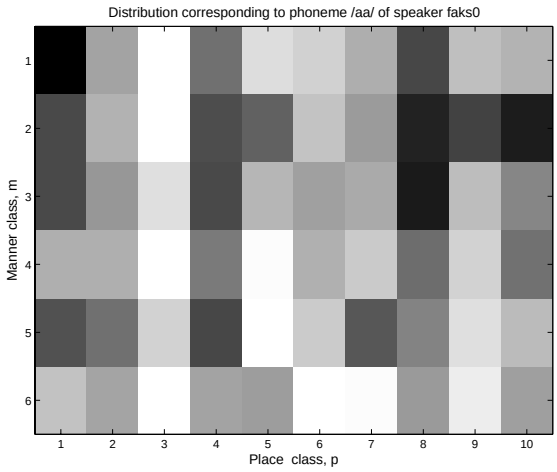
$$\hat{P}_s(m,p\,|\,q) = \hat{P}_s(\text{Manner}=m, \text{Place}=p\,|\,\text{Phoneme}=q)$$
$$= \beta_s^q P_s(\text{Manner}=m, \text{Place}=p\,|\,\text{Phoneme}=q)$$
$$+ (1-\beta_s^q) P_{bg}(\text{Manner}=m, \text{Place}=p\,|\,\text{Phoneme}=q)$$

$\beta_s^q \in [0,1]$ is a phoneme-dependent adaptation coefficient controlling the contribution of the speaker data on the adapted model.

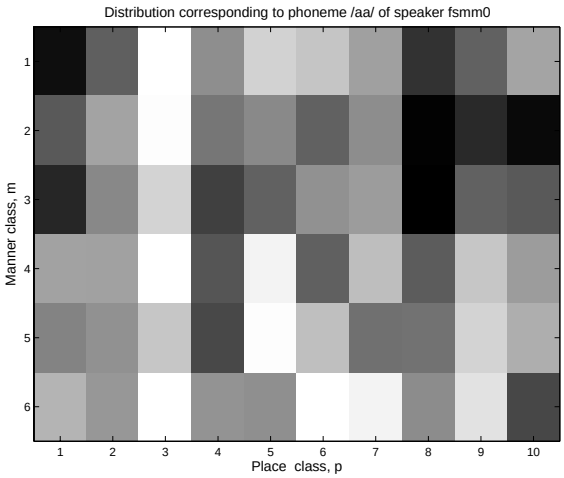
$$\beta_s^q = \frac{\#(\{q\} \text{ in the data of speaker } s}{\#(\{q\}) \text{ in the data of speaker } s + r}$$



Comparing Two Speaker Models



faks0: New England dialect



fsmm0: Southern dialect

Speaker *fsmm0* is acoustically closest to *faks0* among 381 speakers in HTIMIT. These two speakers spoke the same set of sentences (SA sentences).

AFCPM Scoring

- AFCPM Score:

Speaker score

Background score

$$S_{AFCPM} = S_s - S_{bg}$$

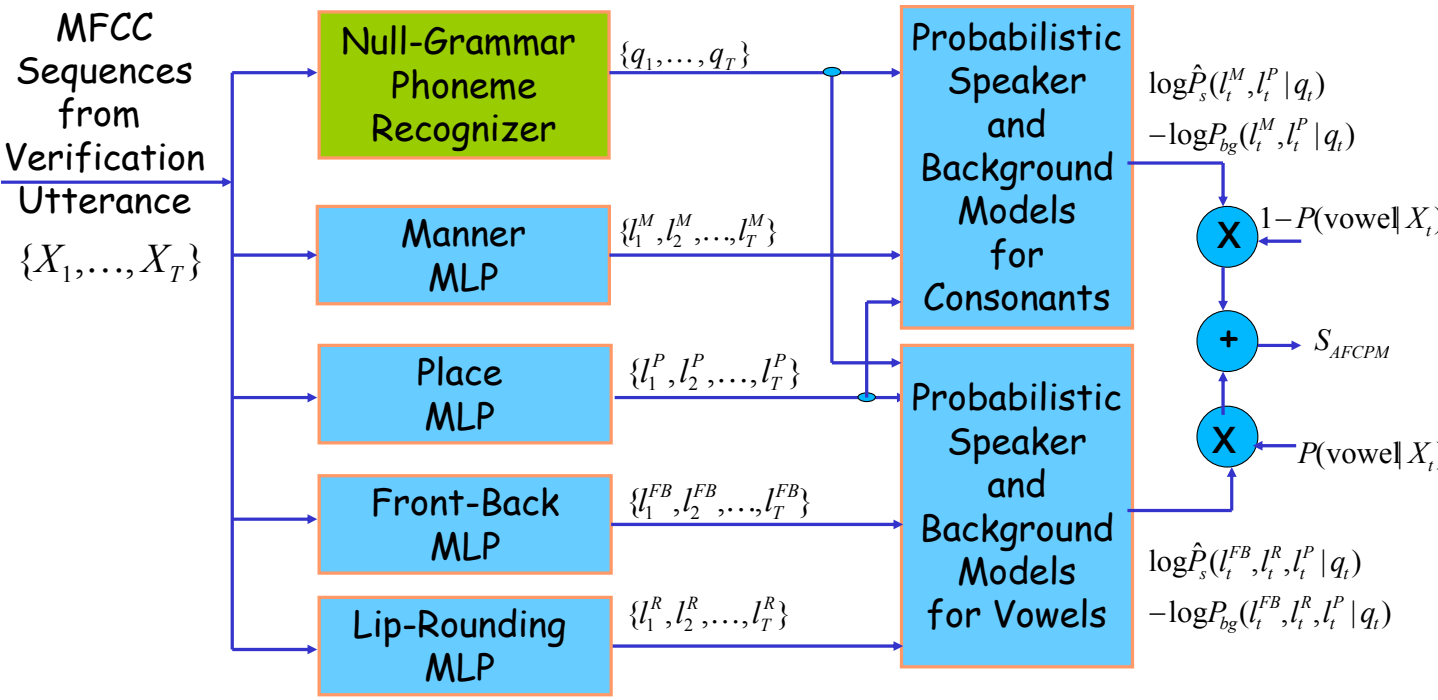
$$= \sum_{\substack{p_s(X_t)\neq 0, p_{bg}(X_t)\neq 0 \\ q_t \neq \text{Silence}}} [\log p_s(X_t) - \log p_{bg}(X_t)]$$

where

$$\log p_s(X_t) = P(\text{vowel}\,|\,X_t)\log \hat{P}_s(l_t^{FB}, l_t^R, l_t^P\,|\,q_t) + [1 - P(\text{vowel}\,|\,X_t)\log \hat{P}_s(l_t^M, l_t^P\,|\,q_t)$$

$$\log p_{bg}(X_t) = P(\text{vowel}\,|\,X_t)\log P_{bg}(l_t^{FB}, l_t^R, l_t^P\,|\,q_t) + [1 - P(\text{vowel}\,|\,X_t)\log P_{bg}(l_t^M, l_t^P\,|\,q_t)$$

AFCPM Scoring

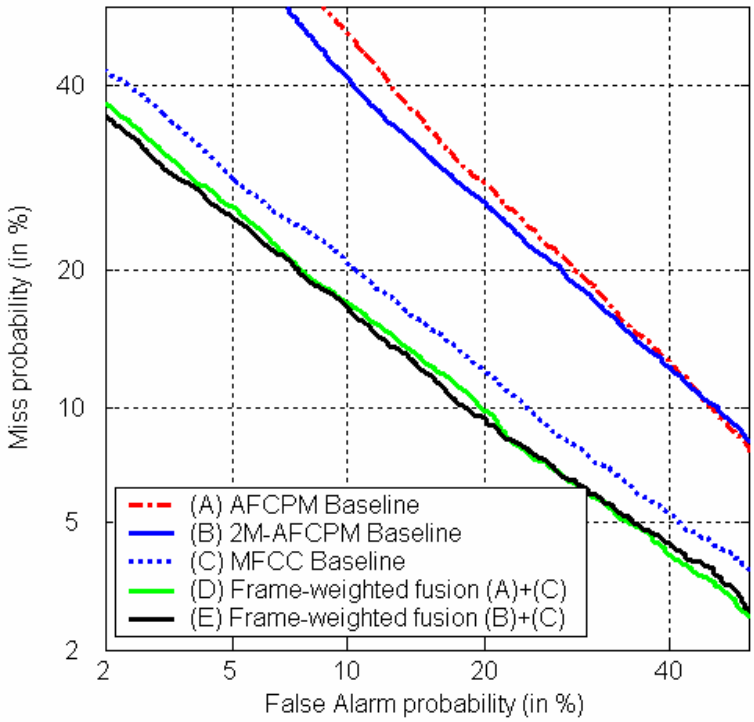


Experiments

- SPIDRE corpus (subset of the Switchboard corpus)
  - 44 target speakers, each with 5 minutes of enrollment speech
  - 200 impostor utterances from 160 non-target speakers. Each utterance is splitted into speaker turns for verification and the duration of each segment is around 1-15 seconds long.
- AFCPM system
  - 46 context-independent phoneme models (3-state 16-mix HMMs) for the null-grammar recognizer.
  - 4 AF-MLPs trained from 3794 utterances selected from the HTIMIT corpus.
- MFCC system
  - A universal background model with 512-mixtures (12-D MFCC + delta).
  - Speaker models: adapted from the background model via MAP.

Features	EER (%)		
	Matched	*Mismatch*	*All*
MFCC	7.59	18.08	15.29
AFCPM	18.07	26.69	24.04
2M-AFCPM	16.69	25.91	23.23
MFCC + AFCPM	7.09	16.31	13.77
(error red. %)	(6.59)	(9.78)	(9.94)
MFCC + 2M-AFCPM	7.15	15.68	13.34
(error red. %)	(5.79)	(13.27)	(12.75)

Table 2: EERs and relative error reduction (in %) obtained by the MFCC system, the AFCPM systems, and the fusion of the two systems. *AFCPM* denotes the MAP-adapted AFCPM system [5]. *2M-AFCPM* denotes the MAP-adapted AFCPM system with two-mixture models proposed in this paper. *MFCC + AFCPM* (*MFCC + 2M-AFCPM*) denotes the fusion of frame-weighted MFCC scores and AFCPM (AFCPM with two-mixture models) scores according to Eq. 10. *Matched* (*Mismatched*) refers to the cases where the handset used by a target speaker in a verification session is identical to (different from) the one used by himself or herself during the enrollment session. The test data from nontarget speakers under *Matched* and *Mismatched* are identical. *All* represents the overall EERs obtained from gathering all test data from the target speakers using both matched and mismatched handsets.



References:

1. K.Y. Leung, M.W. Mak, M.H. Siu, and S.Y. Kung, "Adaptive Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification, *Speech Communication*, 2005.
2. K.Y. Leung, M.W. Mak, M.H. Siu, and S.Y. Kung, "Speaker Verification Using Adapted Articulatory Feature-based Conditional Pronunciation Modeling", *ICASSP'05,* vol. 1, pp. 181-184, 2005,
3. K.Y. Leung, M.W. Mak and S.Y. Kung, "Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification", *ICSLP'05*, Oct. 2004, 2597-2600.