

Beyond pronunciation and fluency: automated evaluation of prosody and accentedness

**LTRC 2014 Amsterdam
June 5, 2014**

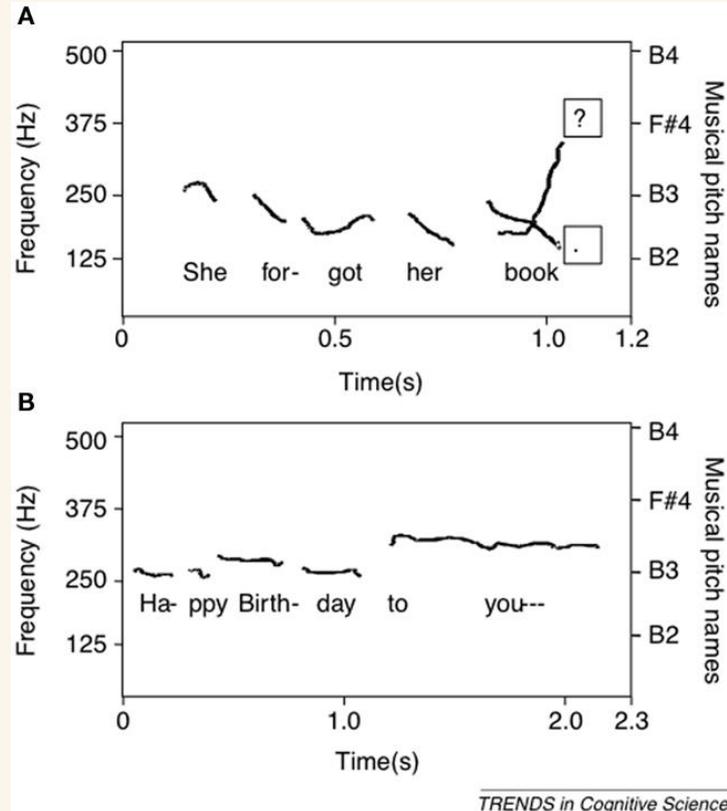
**Jian Cheng
Masa Suzuki
Bill Bonk**

Background

- Automated speech evaluation system in operation for various types of language assessment
 - Proficiency measurement: e.g., TOEFL Practice Online, PTE Academic, Versant, Carnegie Speech
 - Pronunciation feedback system: e.g., EduSpeak
- Commonly scored traits: Pronunciation, Fluency, Vocabulary, Grammar
- Can automated speech evaluation system be trained to evaluate other traits in L2 adult learner speech?

Research Study 1: Automated Prosody Evaluation System

Why study prosody?



TRENDS in Cognitive Sciences

Oral reading fluency as a measure of reading comprehension

- In reading, fluency is the ability to read text aloud quickly, accurately, and with proper expression
- In L1 literacy acquisition literature, oral reading fluency has been shown to be a useful measure of reading comprehension and achievement among school-age children:

Hudson, R. F., Lane, H. B., and Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher* (58), 702– 715.

Shinn, M. R., (2001). Best practices in curriculum-based measurement. *Best practices in school psychology IV*, A. Thomas and J. Grimes, Eds., National Association of School Psychologists, Bethesda, MD.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. *Handbook of Reading Research (Vol. 2)*, R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson, Eds., 418–452.

Research Study 1: Machine-scored prosody

Suzuki, et. al. (2008): Automated method to evaluate rhythm and intonation of sentences read aloud by Japanese learners of English

- It only dealt with short sentences (avge.: 6 words in passage)
- Poor model performance

Maier, et. al. (2009): System to evaluate intonation of German text

- Longer passages (183 words) but for well-rehearsed reading only
- Better model performance

Research Study 1: Prosody Evaluation System

Limitations of these studies

- Conducted in a well-controlled experimental environment
- Typically, short passages
- Systems designed to deal with limited range of L1 background speakers

For a system to be useful for wider assessment use, the system should deal with many L1 backgrounds, and not be dependent on highly controlled settings

Research Study 1: Prosody Evaluation System

Context and Data

- Pearson Test of English Academic (PTE Academic)
- 85 read-aloud passages
- Uses operational automated speech recognition system

Example Passage

Photography's gaze widened during the early years of the twentieth century and, as the snapshot camera became increasingly popular, the making of photographs became increasingly available to a wide cross-section of the public. The British people grew accustomed to, and were hungry for, the photographic image.

Research Study 1: Prosody Evaluation System Rubric for Human Raters



Rating		Capabilities
5	NATIVE Intonation	<ul style="list-style-type: none">• Candidate's intonation is expressive and appropriate throughout the reading, reflecting a clear understanding of the text.
4	ADEQUATE Intonation	<ul style="list-style-type: none">• Candidate's intonation is adequate for the most part, reflecting a good level of understanding of the text. There may be patches of speech with inappropriate intonation or little variation in tone.
3	SOMEWHAT APPROPRIATE Intonation	<ul style="list-style-type: none">• Candidate's intonation is appropriate at times to convey the meaning of the text. There are still relatively long stretches with inappropriate intonation, little variation in tone, or monotonous reading.
2	POOR Intonation	<ul style="list-style-type: none">• Candidate's delivery is monotonous for the most part and shows little variation in tone to reflect the meaning of the text.
1	BAD Intonation	<ul style="list-style-type: none">• Candidate doesn't seem to understand the text. Text is just produced as a string of unrelated lexical items.
0	NO EVIDENCE	<ul style="list-style-type: none">• Silence, or irrelevant material or completely unintelligible material.

Research Study 1: Training an Automated Prosody Evaluation System

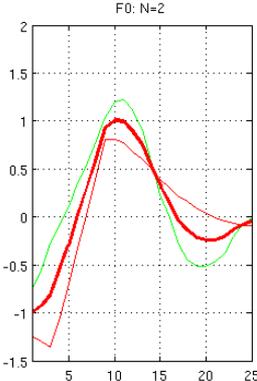
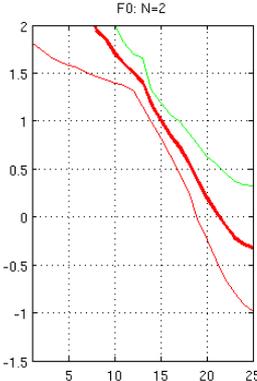
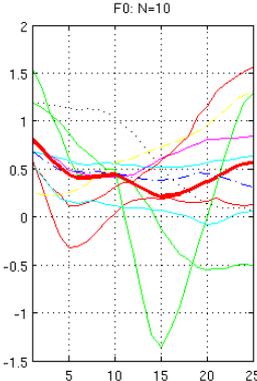
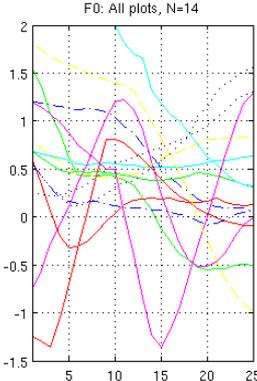
- Training Data
 - 80 adult learners of English per passage
 - 15 speakers of English as first language per passage
 - A separate 340 responses (4 responses per passage) for fine-tuning models
 - Every response was rated by 2 human raters
- Validation Data
 - 158 subjects, randomly selected from a larger pool
 - 357 valid responses, each rated by 4 human raters ($r=0.75$)
 - No data from validation subjects were used during model training

Research Study 1: Prosody Evaluation System

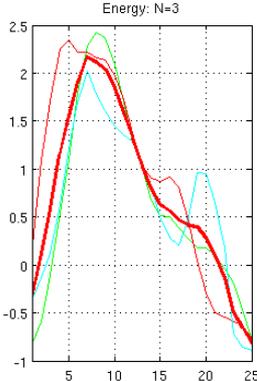
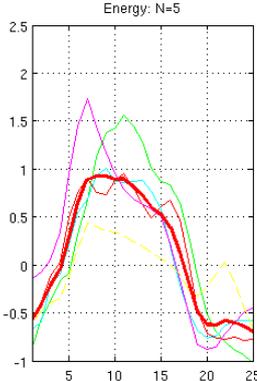
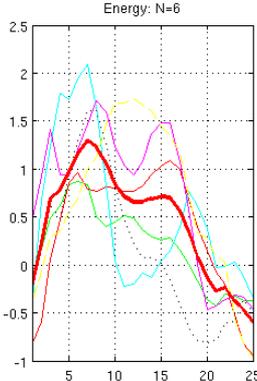
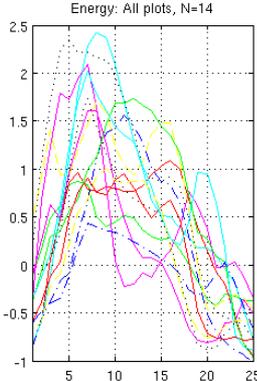
- Intonation and Energy Models
 - Fundamental frequency (F0) contours
 - Energy contours
 - Phoneme durations (log likelihood)
 - Inter-word silence durations (log likelihood)

Example Word: Strategy

F0 Contours



Energy Contours



Research Study 1: Prosody Evaluation System

- Validation results by feature sets

Features	Correlation
F0 Contours	0.67
Energy	0.67
F0 + Energy	0.73
Log interword silence duration probability	0.54
Log phoneme segment duration probability	0.76
Linear regression with all variables	0.80

Research Study 1: Prosody Evaluation System

- Using F0, energy and duration statistics, machine-produced prosody scores correlated quite highly with human prosody ratings ($r = 0.80$)
- This correlation was even higher than the inter-rater reliability correlation between human raters ($r = 0.75$)
- Machine learning techniques can handily implement an assessment of prosody as defined
- This approach needs to be validated with actual comprehension data

Research Study 2: Automated Accent Quantification System

Research Study 2: Rationale

- In call centers and BPOs, an increased demand to be able to detect the heaviness of an accent for job assignment, or give them additional training to refine their accents as appropriate for their jobs
- In L2 performance, degree of accent familiarity affects intelligibility (Ockey, 2014). Accentedness is therefore a relevant construct for the assessment of speaking in the context of the perceived value of particular speech varieties

Research Study 2: Motivation

- RQ 1: Is it possible to develop an automated system to classify speakers of English according to their degree of “Indian accentedness” as judged by a group of raters?
- RQ 2: Do results correlate highly with ratings assigned by human raters for a validation dataset?

Research Study 2: Characteristics of Indian English Accents

- Indian varieties of English tend to be syllable-timed as opposed to stress-timed
- Indian English tends to have a reduced vowel system compared with North American or British English
- Indian English is typically associated with a different pronunciation of some consonants from that of North American or British English

Research Study 2: Indian Accent

- Trudgill and Hannah (2008) identified 13 phonemes as general features of Indian speakers of English

Consonant Categories	Phonemes
Labiodental fricative	/v/
Bilabial approximant	/w/
Plosives	/p/, /t/, /k/
Alveolars	/t/, /d/, /s/, /z/, /l/, /r/
Postalveolar fricatives	/zh/, /sh/
Postalveolar affricate	/ch/

Research Study 2: Experimental Data

- 825 participants – mix of L1 English, Indian English, and other L2 speakers, both genders
- 825 participants' data were divided into three sets
 - Training dataset ($n=411$)
 - Development dataset ($n=206$)
 - Test dataset ($n=208$)
- Read Aloud passages from PTE Academic and sentences from Versant English Test
- Operational test data
- Average number of words per passage = 50 words
- Candidates had an average of 2.3 valid responses for analysis

Research Study 2: Experimental Data

- 2-3 raters rated each response according to Indian English accentedness rubrics

Scale	Criteria
6	<i>Heavy</i> Indian accent: The speaker has a very thick Indian accent.
5	<i>Medium heavy</i> Indian accent: The speaker sounds Indian and the accent is pronounced.
4	<i>Medium light</i> Indian accent: The speaker sounds Indian and the accent is clear.
3	<i>Light</i> Indian accent: The speaker sounds Indian but the accent is subtle.
2	Non-native accent: The speaker sounds like a learner of English (e.g. Korean, Japanese, Philippine), and is probably not Indian.
1	Native: The speaker sounds native (e.g, US, UK, Australian).
0	Unintelligible/Irrelevant: Silence or unintelligible material.

Research Study 2: Experimental Data

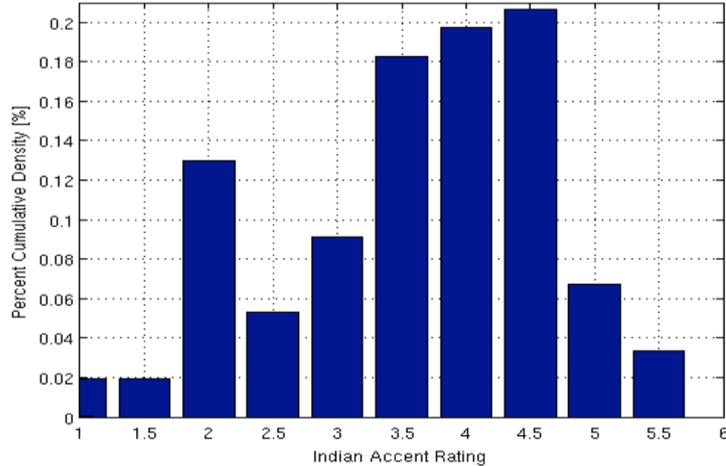


Figure 1: A histogram showing the average Indian accent score distribution for individuals in the test set as rated by humans.

- Average of inter-rater correlations at the response level was 0.774
- Human raters made reasonable judgments about Indian accent

Research Study 2: Predictor Variables

- Four phoneme classes were created as sets of predictor variables

Phoneme Classes	Phonemes
ap	All phonemes
vp	All vowel phonemes
cp	All consonant phonemes
ip	13 phonemes associated with Indian English Speakers

Expect this class to better predict human ratings

- Other features as extracted from speech processing system
 - 2 types of confidence scores extracted from ASR
 - Prosodic features such as phoneme segmental duration and inter-word silence log-likelihoods (as in Study 1)
 - A few spectral likelihood features borrowed from Versant system

Research Study 2: Results

- Prosodic features performed worst in predicting human scores ($r = 0.035-0.057$) at the response level → Excluded from the final model
- Back propagation nonlinear neural net model worked better than multiple linear regression, demonstrating that the problem is nonlinear
- Pearson correlation of 0.84 was achieved between the average of all machine scores and average of all human ratings at the test-taker level
- Indian English phoneme class alone had a correlation of 0.73 – the best predictor variable set as expected

Summary of conclusions

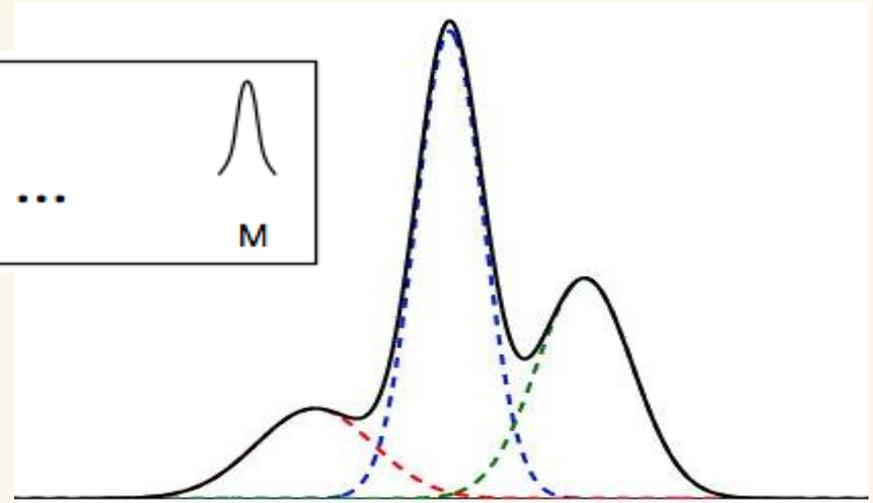
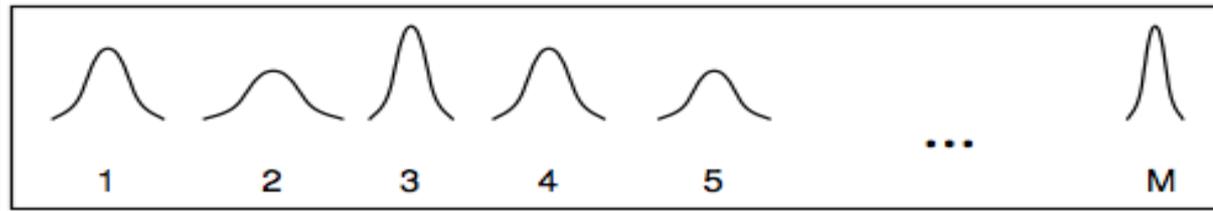
- Traits such as prosody and accent quantification can be automatically evaluated with a reasonable degree of correspondence with human ratings
- We proposed the idea of using GMMs to model only certain phonemes that may have better predictive power in quantifying an Indian accent
- We verified computationally that Indian English has more distinctive features in consonants than in vowels, and that certain consonants have more discriminative power than others
- Prosodic features may not be as useful as phonetic features to quantify an accent
- Accent quantification can be effectively implemented with only 2.3 items administered per candidate
- Next step is to determine how much unique and appropriate information these new measures bring to L2 score estimates

Questions?

Research Study 2: Gaussian Mixture Model

- A GMM is composed of a finite mixture of multivariate Gaussian components:

$$P(\mathbf{x}|M) = \sum_{k=1}^K w_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$



GMM Model Training and Log-Likelihood

- Using all the training data, we built a UBM from the full set of feature vectors of interest.
- We then trained the accent heaviness dependent models by adapting the UBM using the training data from the specified groups via a MAP adaptation procedure. Only mean vector adaptation was performed.

Some Other Features - Prosodic Features

- Energy, pitch and duration.
- The duration statistics models were built from native data from the Versant English Test.
- The statistics of the phoneme durations of native responses were stored as non-parametric cumulative density functions (CDFs).
- Duration statistics from native speakers were used to compute the log likelihood for durations of phonemes produced by candidates. If enough samples for a phoneme in a specific word existed, we built a unique duration model for this phoneme in context.

$$\log_seg_prob = \frac{1}{N-2} \sum_{i=2}^{N-1} \log(\Pr(D_i))$$

Some Other Features - Spectral modeling

- We computed few spectral likelihood features according to native and learner segment models applied to the recognition alignment of segmental units.
- We did force alignment of the utterance on the word string from the recognized sentence using the native mono acoustic model.
- For every phoneme, using the previous time boundary constrain from the native mono acoustic model, we did an allophone recognition using the native mono acoustic model again.
- Different features by using different interested phonemes.

$$spectra_1 = \frac{1}{N} \sum_{i=1}^N \frac{lp_i^{fa} - lp_i^{ap}}{d_i}$$

- ppm: the percentage of phonemes from the allophone recognition matching to the phonemes from the force alignment.

Some Other Features - Confidence modeling

- After finishing speech recognition, we can assign speech confidence scores to words and phonemes. Then for every response, we may compute the average confidence, the percentage of words or phonemes whose confidences are lower than a threshold value as features.

Final Models and Performance Measures

- When developing different GMM models, overfitting to the training data is often unavoidable.
- The models were built using the training data and then tested with the development set.
- For the final model, we used the optimal parameters and combined the training set and the development set for model training. The results were then reported on the test set. The test set was never used to train models.
- PKT tried both simple multiple linear regression models and back propagation neural network models using the log-posterior probabilities in six speaker group.
- We compared Pearson correlation coefficients between machine scores and human ratings.

Experimental Data

- We used recordings of speakers in real assessment environments as they read aloud passages from a high-stakes English test -- Pearson Test of English Academic and from Versant English.
- The average number of words per passage was about 50.
- The sample rate for the recordings was 8 kHz with 8 bits (telephone band).
- We asked human raters to rate the responses according to the rating criteria. Two to three different human raters rated every response.
- Human raters identified responses that had silence, or irrelevant or completely unintelligible material. These responses were excluded from our study.
- On average, every subject provided about 2.3 valid responses.

Experimental Data

Table 3: The total speech segment duration in different sets for different phoneme classes, in hours. n is the number of subjects in the corresponding sets.

Phoneme class	<i>ap</i>	<i>vp</i>	<i>cp</i>	<i>ip</i>
Training ($n = 411$)	4.44	1.67	2.77	1.80
Development ($n = 206$)	2.21	0.83	1.38	0.90
Test ($n = 208$)	2.28	0.85	1.43	0.93

- The average of the inter-rater correlations at the response level was 0.774.
- This level of correlation indicates that the human raters made reasonable judgments about Indian accent.

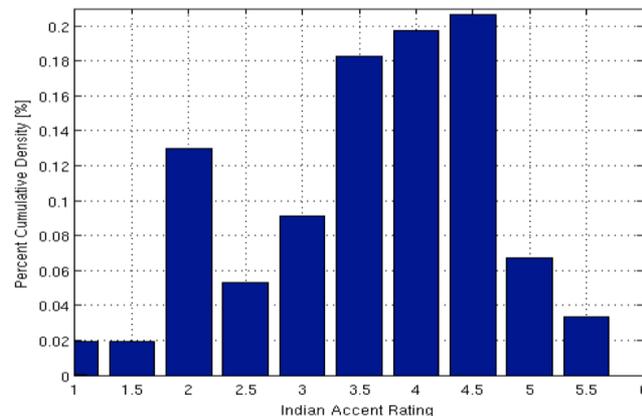
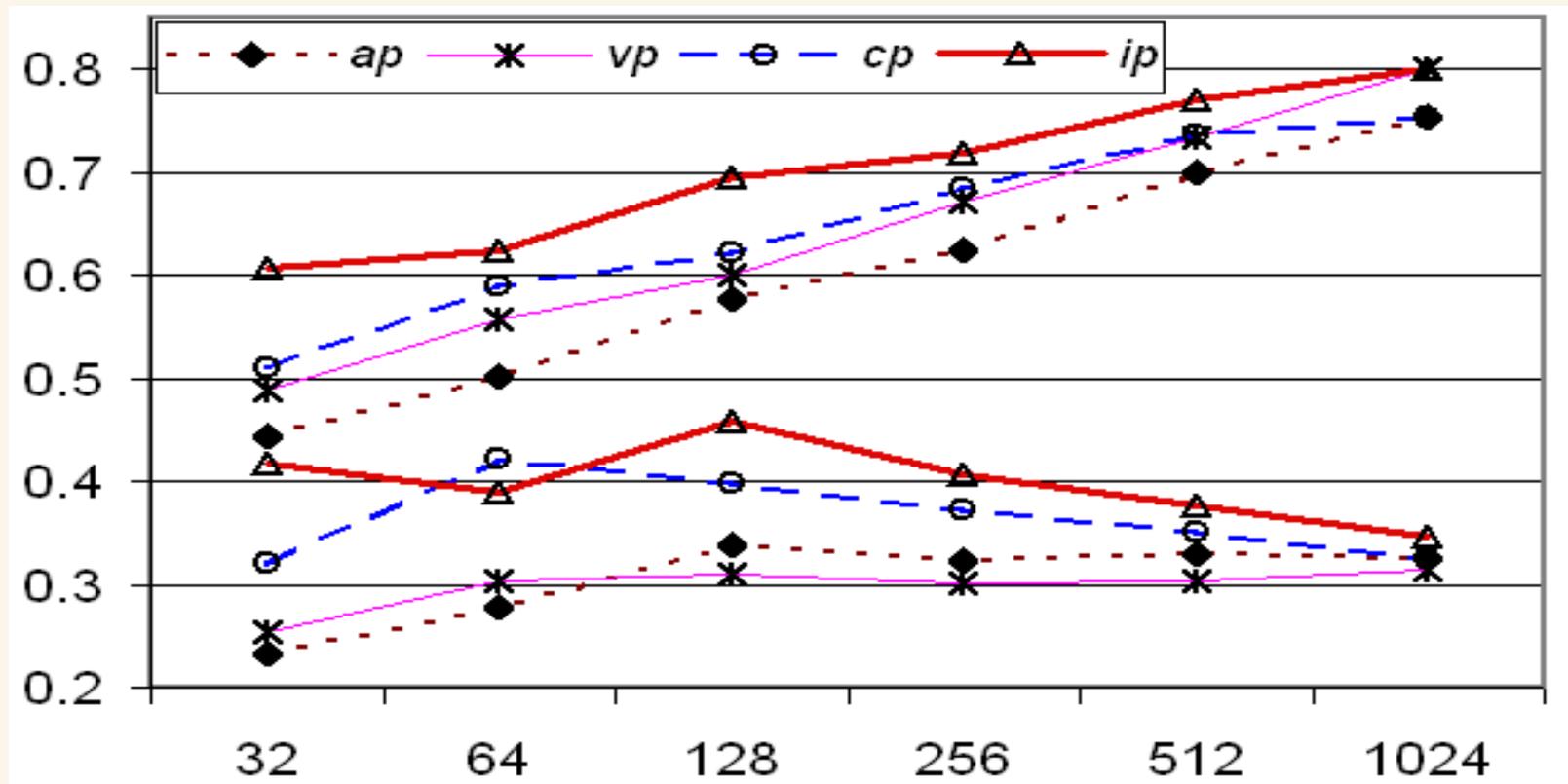
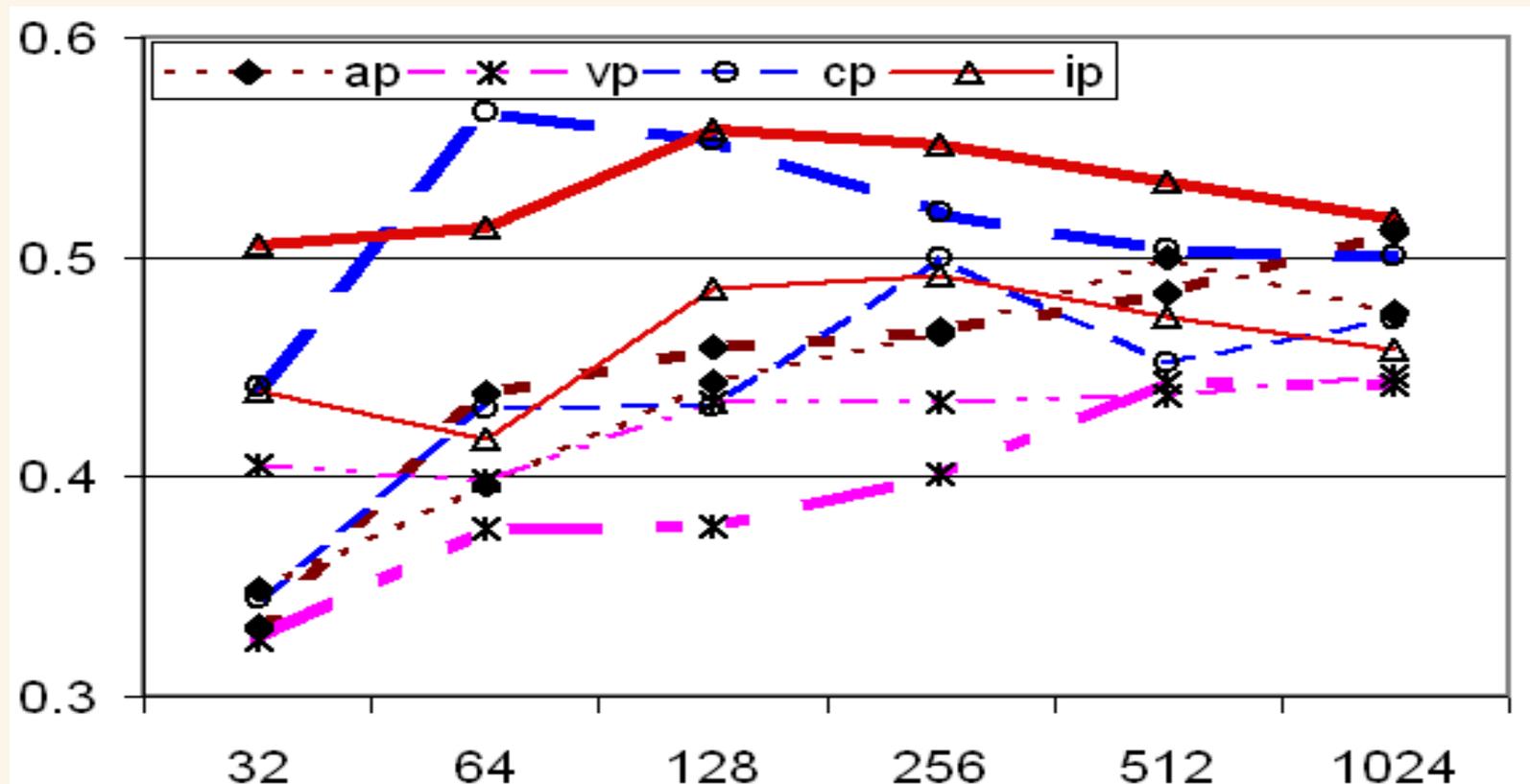


Figure 1: A histogram showing the average Indian accent score distribution for individuals in the test set as rated by humans.

Experimental results – GMM Parameters (LR)



Experimental results – GMM Parameters (NN)



Correlations at the response level using different features in the development set

Features	Correlation
<i>ip</i> ₁₋₆	0.558
<i>acon</i> f^μ	0.234
<i>lcon</i> f^μ	0.184
<i>log_seg_prob</i>	0.035
<i>iw_log_seg_prob</i>	0.057
<i>ip</i> ₁₋₆ , <i>cp</i> ₁ , <i>cp</i> ₃ , <i>cp</i> ₆ , <i>ap</i> ₁₋₄ , <i>vp</i> ₁ , <i>vp</i> ₄₋₅	0.610
the final model	0.674

- The average of the inter-rater correlations at the response level in the development set was 0.739.

Correlations using different features in the test set

Features	Correlation
ip_{1-6}	0.733
$ip_{1-6}, cp_1, cp_3, cp_6, ap_{1-4}, vp_1, vp_{4-5}$	0.741
the final model	0.779

- The average of the inter-rater correlations at the response level in the test set was 0.774.
- If we use the average of all human ratings as the participant's final human score and the average of all machine scores as the participant's final machine score, at the participant level, the final correlation was 0.84. This result was achieved by using only about 2.3 read-aloud items.

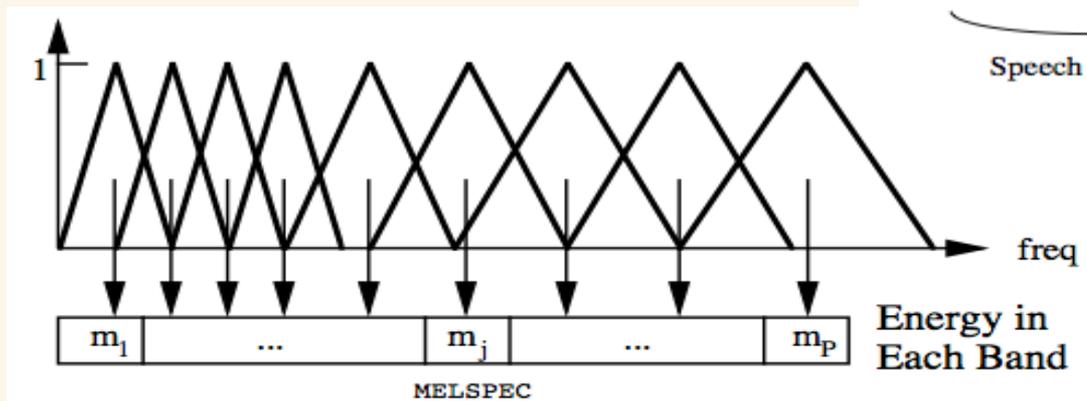
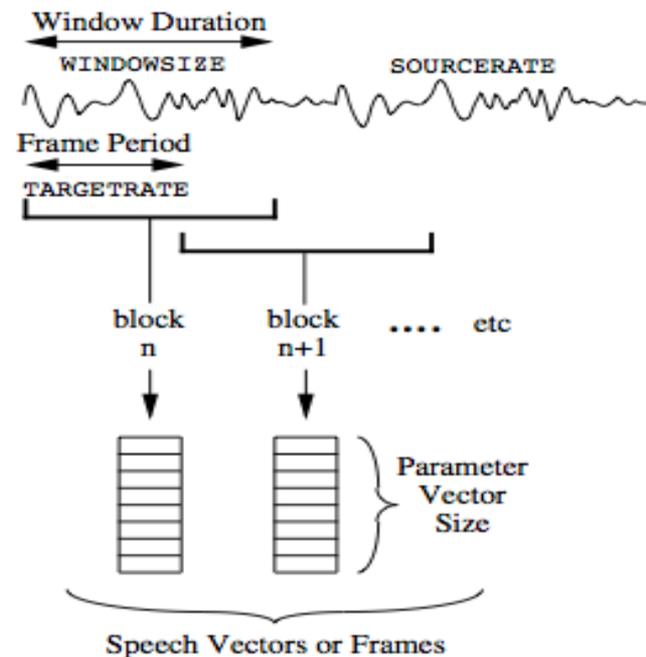
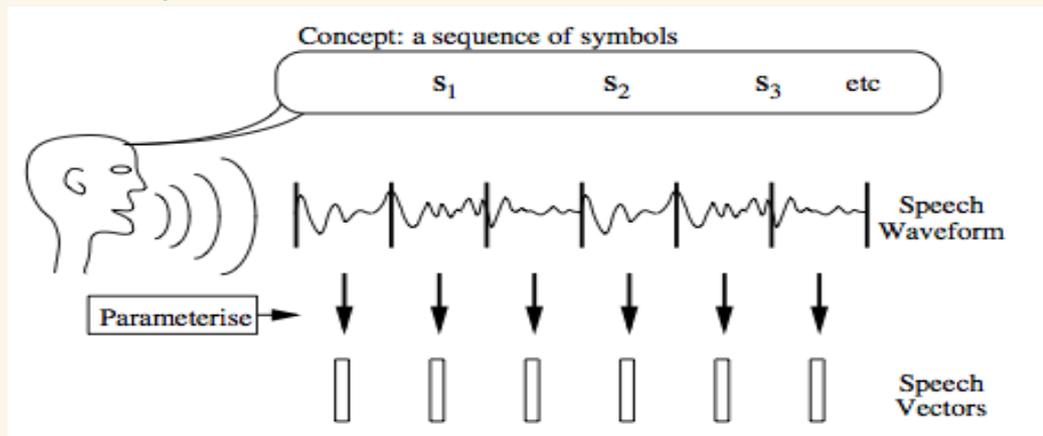
Discussion

- The GMM models used here are gender-independent. We expect that gender-dependent models may perform better as gender-dependent models were trained frequently in accent classification tasks.
- Compared to the performance of GMM models that were trained using the training set, the significant performance improvement observed when using both the training and development sets reveals that collecting more data may be able to help improve performance.
- When we have enough data, we may increase the number of GMM components to further improve the performance.

Conclusions

- We used GMMs successfully for modeling accent spectral characteristics in different groups of subjects.
- We proposed the idea of using GMMs to model only certain phonemes that may have better predictive power in quantifying an Indian accent.
- We verified computationally that Indian English has more distinctive features in consonants than in vowels, and that certain consonants have more discriminative power than others.
- We concluded that prosodic features may not help to quantify an accent.
- We achieved a human-machine correlation coefficient of 0.78 at the response level and 0.84 at the participant level.
- The results support our hypothesis that our new proposed methods can successfully quantify an accent automatically.

GMM Input Features



Gaussian Mixture Model

- After we extracted interested feature vectors from a recording:

- $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ averaged log-likelihood is defined as:

$$LL_{avg}(\mathbf{X}, M_i) = \frac{1}{T} \sum_{t=1}^T \log P(\mathbf{x}_t | M_i)$$

- One Universal Background Model (UBM) and six other models for each of the six groups of speakers.

- We are more interested in the $P(\mathbf{x} | M_i)$ ability instead of the likelihood $P(M_i | \mathbf{x})$, some simplifications can give:

$$\log P(M_i | \mathbf{x}) \approx LL(\mathbf{X}, M_i) - LL(\mathbf{X}, M_{UBM})$$

- For each utterance, we produced the log-posterior probability in each speaker group model and treated these probabilities as input features for further machine learning.