Computer-Assisted Language Learning (CALL) Systems

Tatsuya Kawahara (Kyoto University, Japan)

Nobuaki Minematsu (University of Tokyo, Japan)

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)

(Traditional) LL → CALL

- (Traditional) LL: magnetic audio tape
  - Single media, Sequential access

- Computer-Assisted LL
  - Multi-media, Random access
    - Easier comparison of learner's speech and model speech
  - Speech technology can be incorporated
    - Partly replace rater's or teacher's jobs

Speech Technology for LL

- Automate assessment of proficiency
  - PhonePass → Versant
  - ETS-TOEFL
  - PSC (Putonghua Shuiping Ceshi)

- Assist LL
  - With light supervision…CALL classroom
  - Self-learning
    - Need to keep motivating…Edutainment
    - Need to avoid enhancement of errors

Target Population of CALL

- Non-native speakers
  - Particular L1 (ex.) English LL for Japanese people
    - Still diverse in proficiency level, but L1 knowledge useful
  - Unlimited L1 (ex.) Japanese LL for people in the world
- Children (native) [Russel 1996]
- Handicapped (Hearing or Articulation-impaired) people [Bernstein 1977]
- Accented (dialect) people
  - Putonghua [Hu 2008]
  - Operators at Call Centers

Target Skill of CALL

- Reading
- Writing
- Listening
- Speaking-Pronunciation
  - Phone, word
  - Sentence, paragraph
  - Segmental, prosodic

- Vocabulary, Grammar
- Pragmatic Dialog (Communication)
  - travel-shopping, business-negotiation

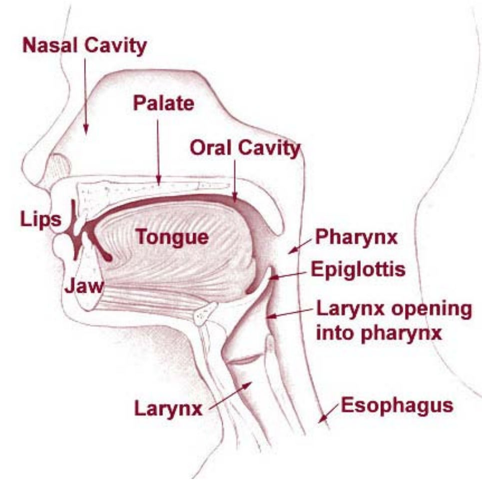
Importance of Pronunciation Training [Bernstein 2003]

*Comm = pron * lex * (1+syn+rhet+prag+soc)*

- comm. = communication skills
- pron. = pronunciation
- lex. = lexical control and vocabulary
- syn. = syntax
- rhet. = rhetorical form
- prag. = pragmatics
- soc. = sociolinguistics

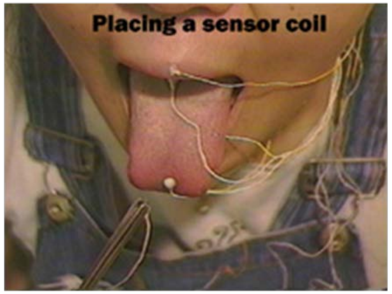
- Pronunciation skill affects entire communicative performance
- Native-sounding pronunciation may not be needed, but acceptable (intelligible enough) pronunciation is desired for smooth communication

Articulation → Speech

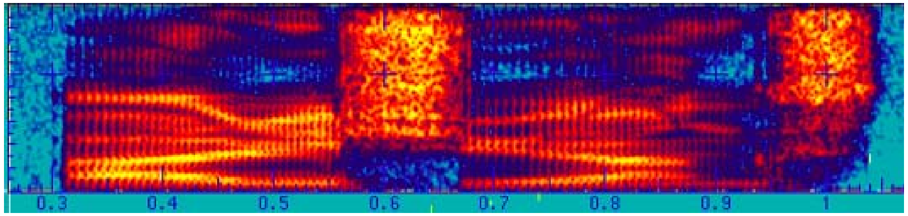


- Students must learn how to control articulators (vocal tract)

- But it is not easy to observe the movement of these organs



- Observation is feasible for acoustic aspect of speech



Visual Presentation of Articulation

- Talking Head showing correct articulation [Massaro 2006]
- Acoustic-to-articulatory inversion to estimate the articulatory movements [Badin 2010]

Segmental and Prosodic Aspects

- Segmental Pronunciation → Kawahara
  - Phonemes (Sub-words)
  - Features: spectrum envelop-based

- Prosody → Minematsu
  - Tones
  - Lexical accents
  - Intonation and rhythm patterns
  - Features: fundamental frequency, power, and duration

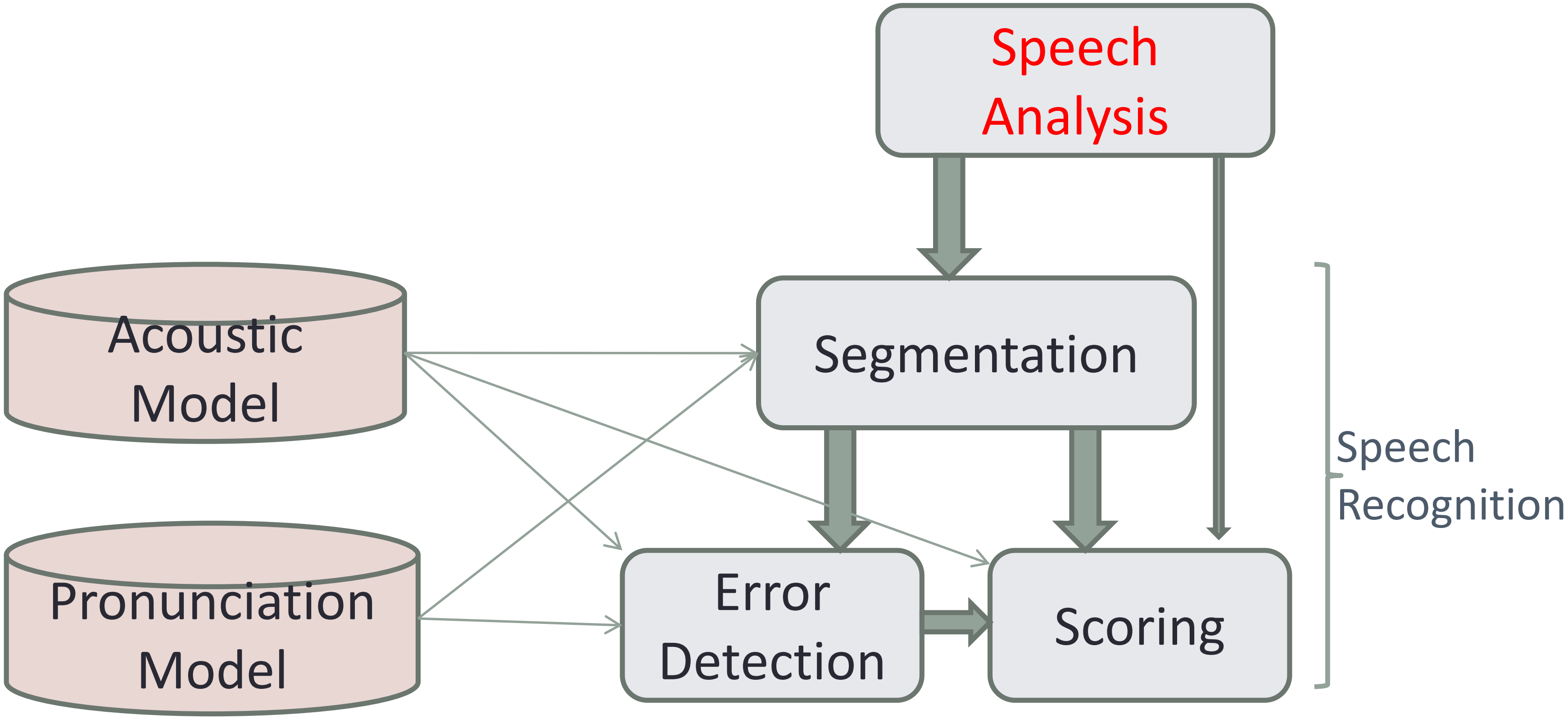
OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)

Speech Technology used in CALL

- Speech analysis
  - spectrum, pitch, power
  - Feature normalization required for objective comparison with model speaker
- (Constrained) speech recognition (ASR)
  - Speech segmentation-alignment
  - Error detection
  - Scoring
  - Need to model non-native speech and handle erroneous input
  - Not only segmental aspect, but also prosodic aspects
- Speech synthesis (Minematsu)

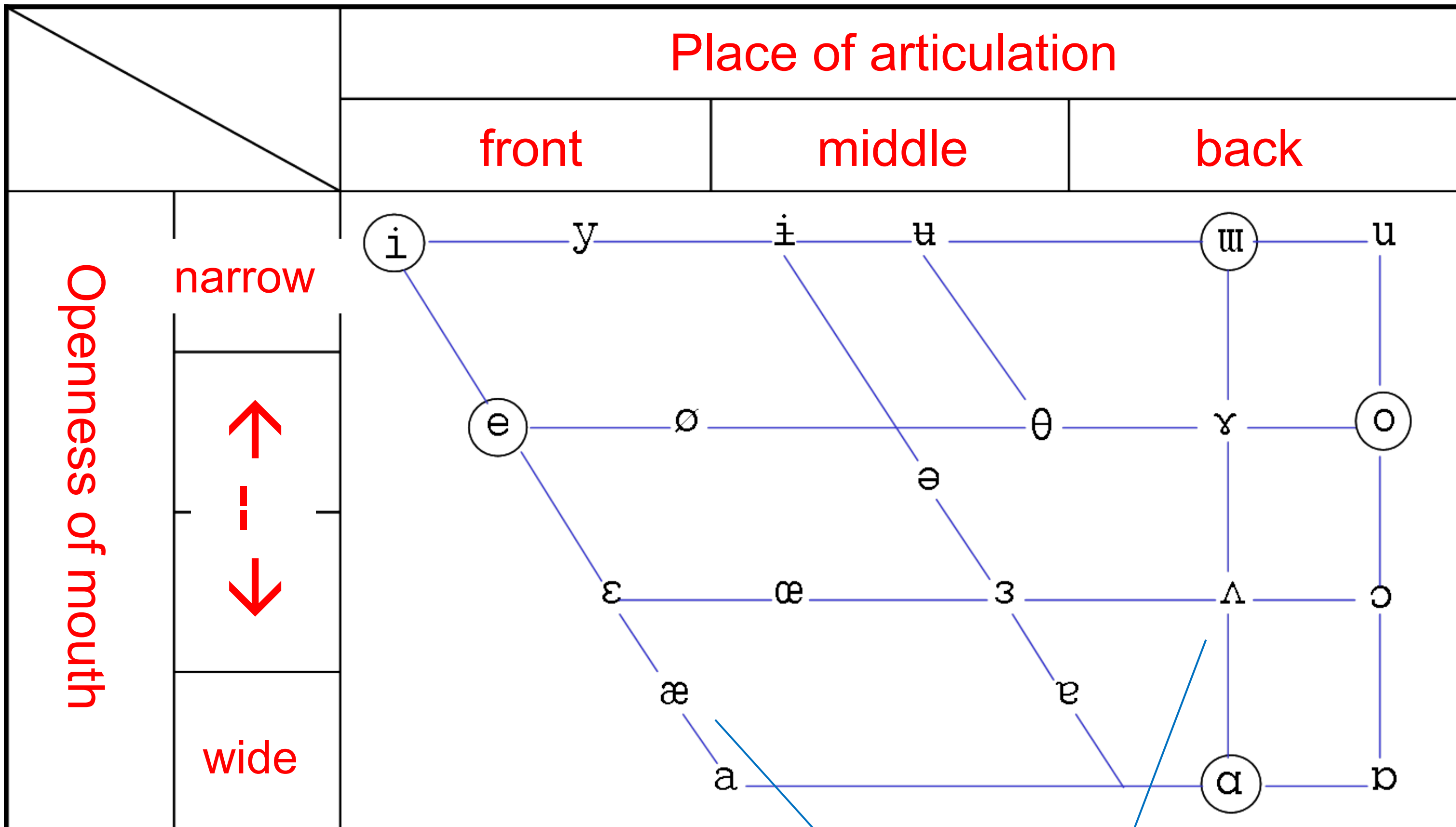
Flowchart of Pronunciation Error Detection and Scoring



Formant and Articulatory Features

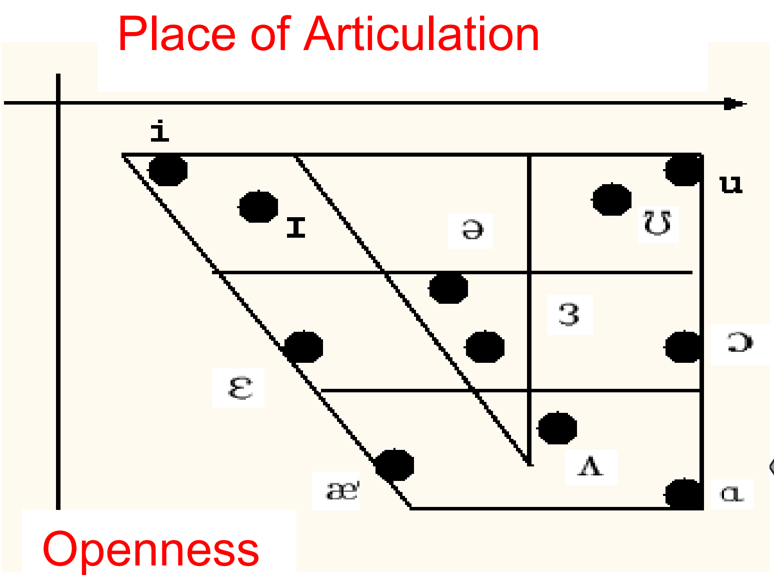
- Potentially useful for effective diagnosis and feedback
  - Direct relationship with articulation
- Not easy to make reliable and robust estimation
  - Not used in ASR

Classification of Vowels

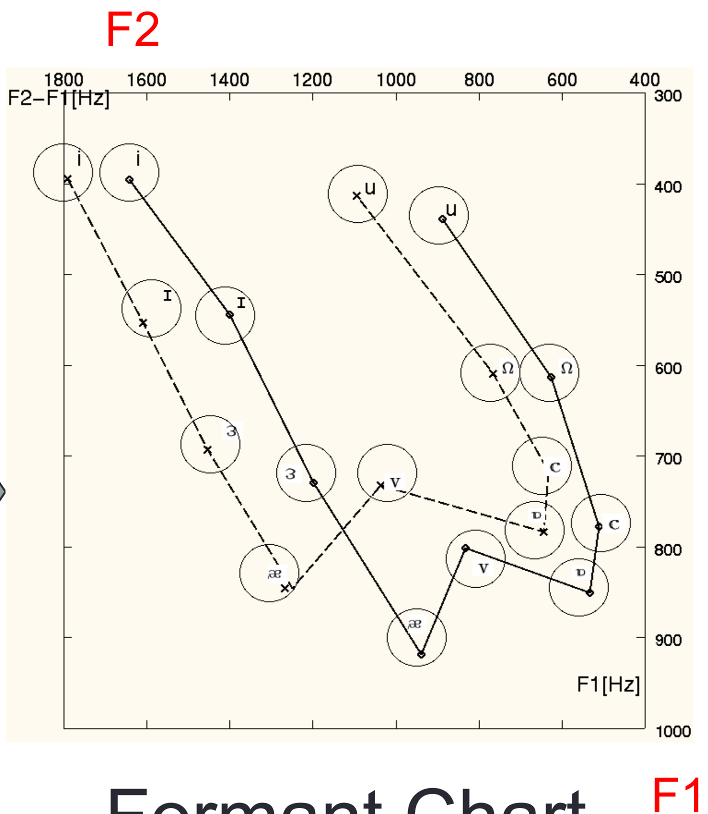


“bat” vs. “but”

Relationship between Articulation and Formants



Articulation Chart



Formant Chart

Classification of Consonants (Japanese)

	Bilabial		Alveolar		Palatal		Glottal
	voiced	unvoiced	voiced	unvoiced	voiced	unvoiced	unvoiced
Fricative		f*)	z	s	ʒ	ʃ	h*)
Affricate			dz	ts	dʒ	tʃ	
Stop	b	p	d	t	g	k	
Semi-vowel	w		r**)		j		
Nasal	m		n		ŋ		

"sea" vs. "she"

MFCC: Mel-Frequency Cepstrum Coefficient

- Most widely-used spectral feature
  - Mel-bandwidth ← human perception
  - Cepstrum → spectrum envelope
    - orthogonal & less correlated → appropriate for statistical model

1. DFT(FFT) → power spectrum
2. Mel-conversion (Mel-band filter bank)
3. Logarithm + Cosine Transform (IDFT) → cepstrum
4. Extract low quefrency (12) coefficients

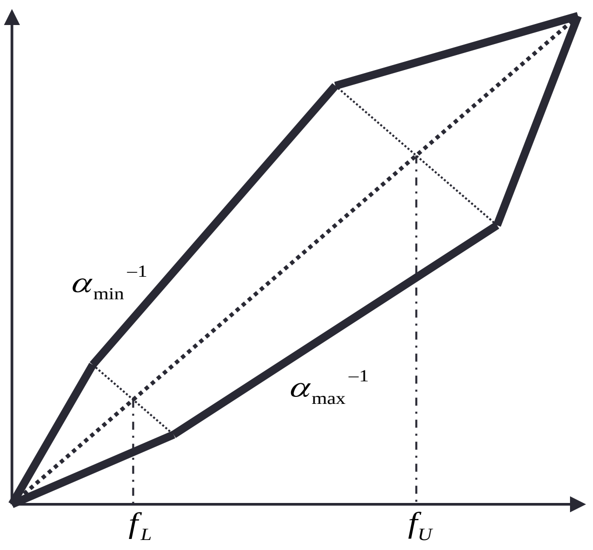
Feature Normalization in Speech Analysis

- Feature normalization
  - for objective comparison with model speaker
  - for score calculation via speech recognition
  - against speakers (native/non-native)
  - against acoustic channels (database/users)

- Normalization methods for MFCC
  - Cepstrum Mean Normalization (CMN)
  - Cepstrum Variance Normalization (CVN)
  - Histogram Equalization

Speaker Normalization in Speech Analysis

- Vocal-Tract Length Normalization (VTLN)
  - Warping spectral dimension
  - Based on acoustic model likelihood



- Pronunciation Structure (by Minematsu)
  - Invariant-feature (F-divergence)

Speech Recognition for CALL

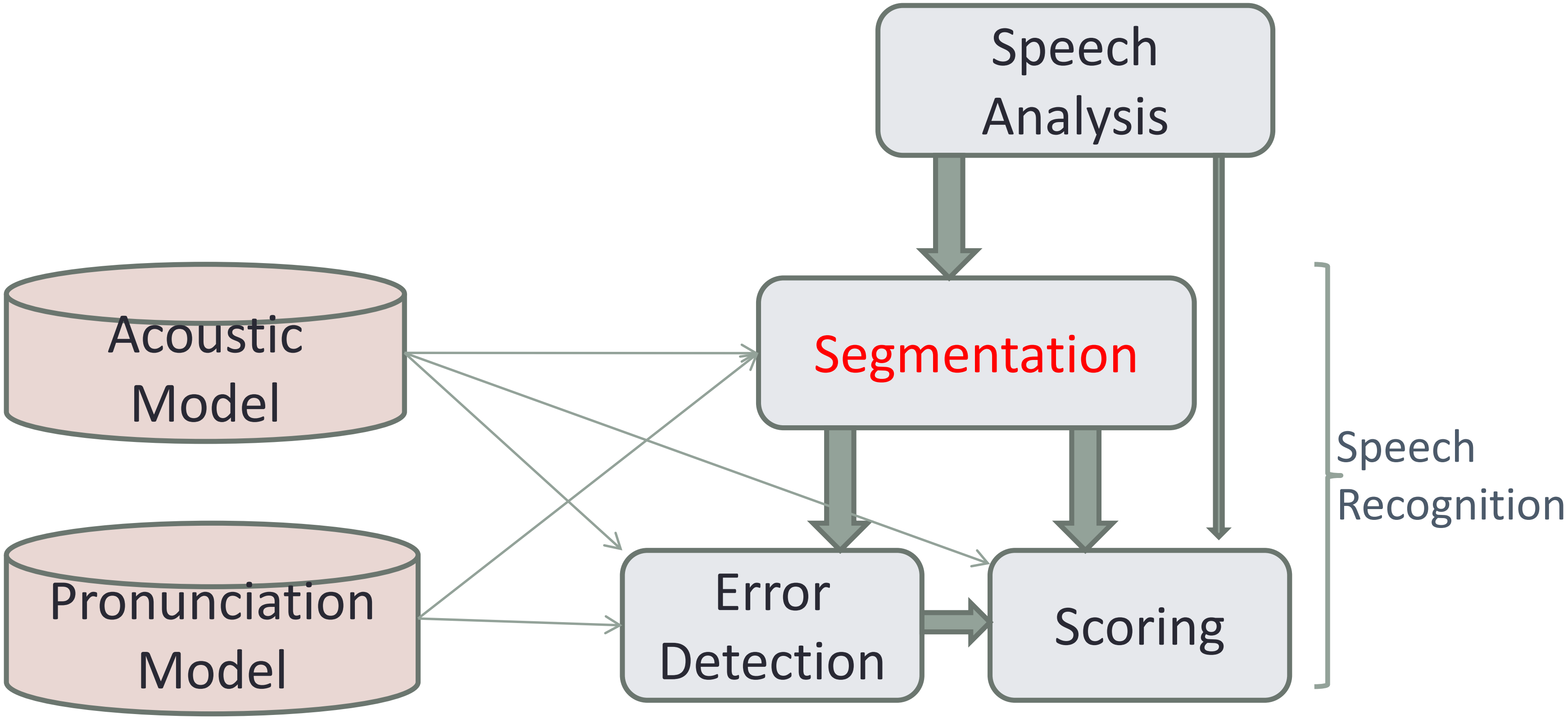
- Tasks
  - Speech segmentation-alignment
  - Error detection
  - Scoring
- Challenges
  - Modeling non-native speech
  - Handling erroneous speech input
- Constraint
  - Target word or sentence is given

ASR vs. CALL

X: speech input, W: phone label ← word sequence (target)

- ASR
  - For given X, find W that maximizes $p(W|X)$
  - Solved by max $p(W)*p(X|W)$
  - Each phone model $p(x|w)$ is trained
- CALL
  - W (oracle) and X (not reliable) given,
  - Segmentation: Viterbi forced alignment
  - Error detection: find W' such that $p(X|W')>p(X|W)$
  - Scoring: evaluate $p(X|W)$?? How to train the model??

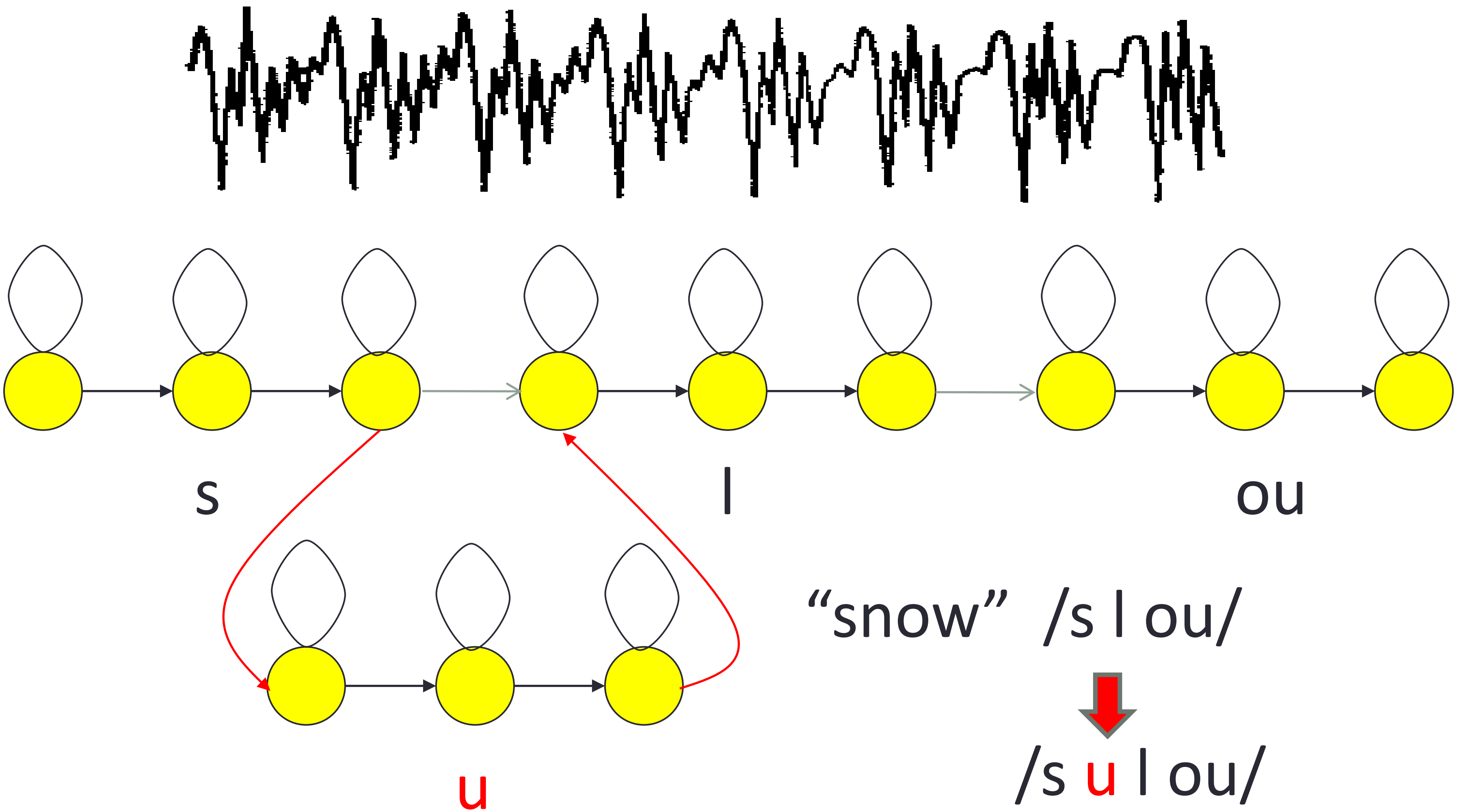
Flowchart of Pronunciation Error Detection and Scoring



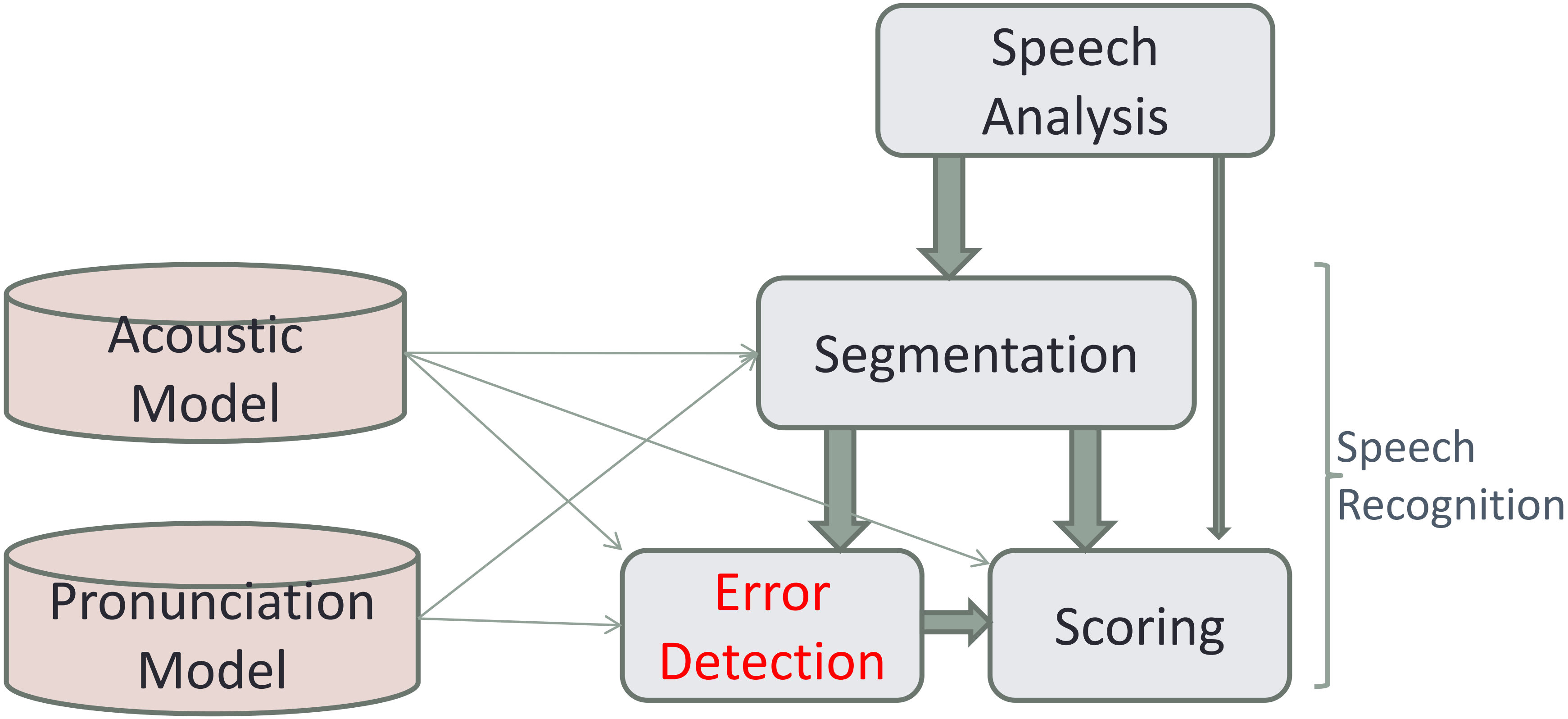
Segmentation

- Pre-process for scoring
- Viterbi forced alignment with HMM representing W
- In fact, there may be pronunciation errors in X
  - Insertion & deletion seriously affect alignment
  - Error prediction/detection may be necessary

Segmentation



Flowchart of Pronunciation Error Detection and Scoring

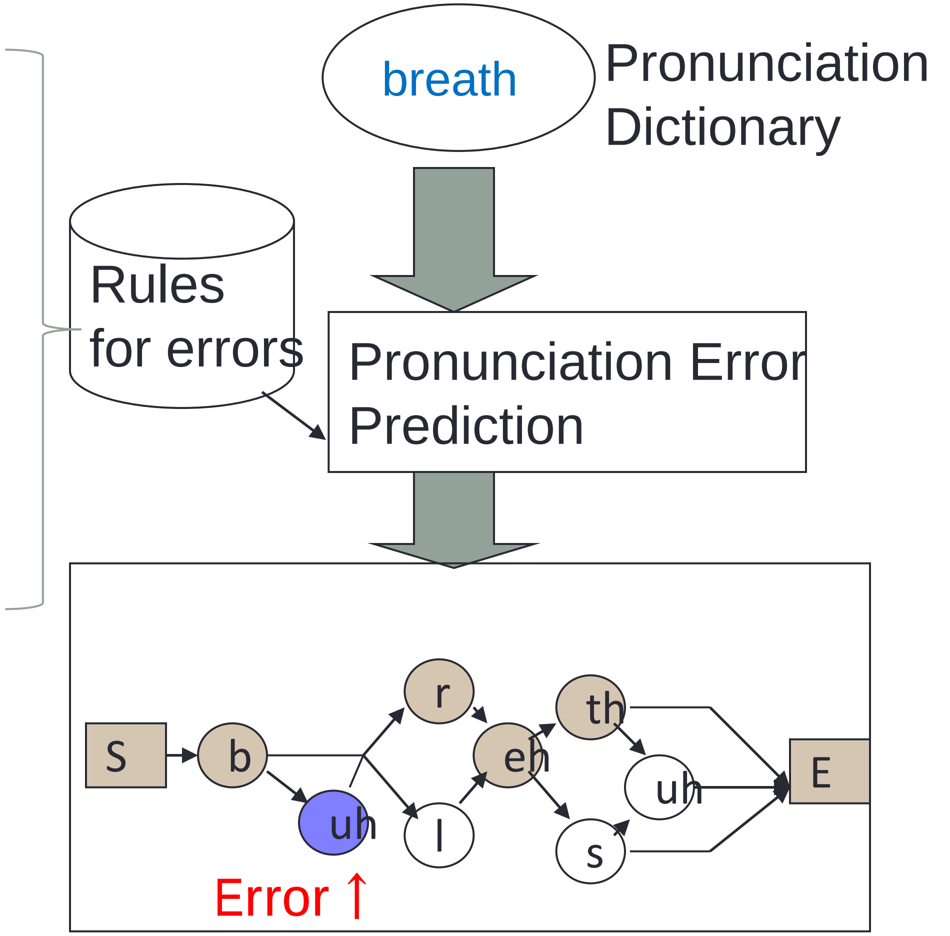


Error Detection

- Find W' such that $p(X|W') > p(X|W)$
- Compute scores $p(x|w')$ for alternative phones w' for each segmented region x
- When we take into account insertions and deletions, we need to generate a network of possible errors
- Error prediction can be done with prior knowledge, such as L1
  - Alternative phones w' can be taken from L1

Error Prediction in Pronunciation Model

- No equivalent syllable in L1
  (ex.) sea → she
- No equivalent phoneme in L1
  (ex.) l → r, v → b
- Vowel insertions
  (ex.) b-r → b-uh-r



Error Detection based on Classification Approach

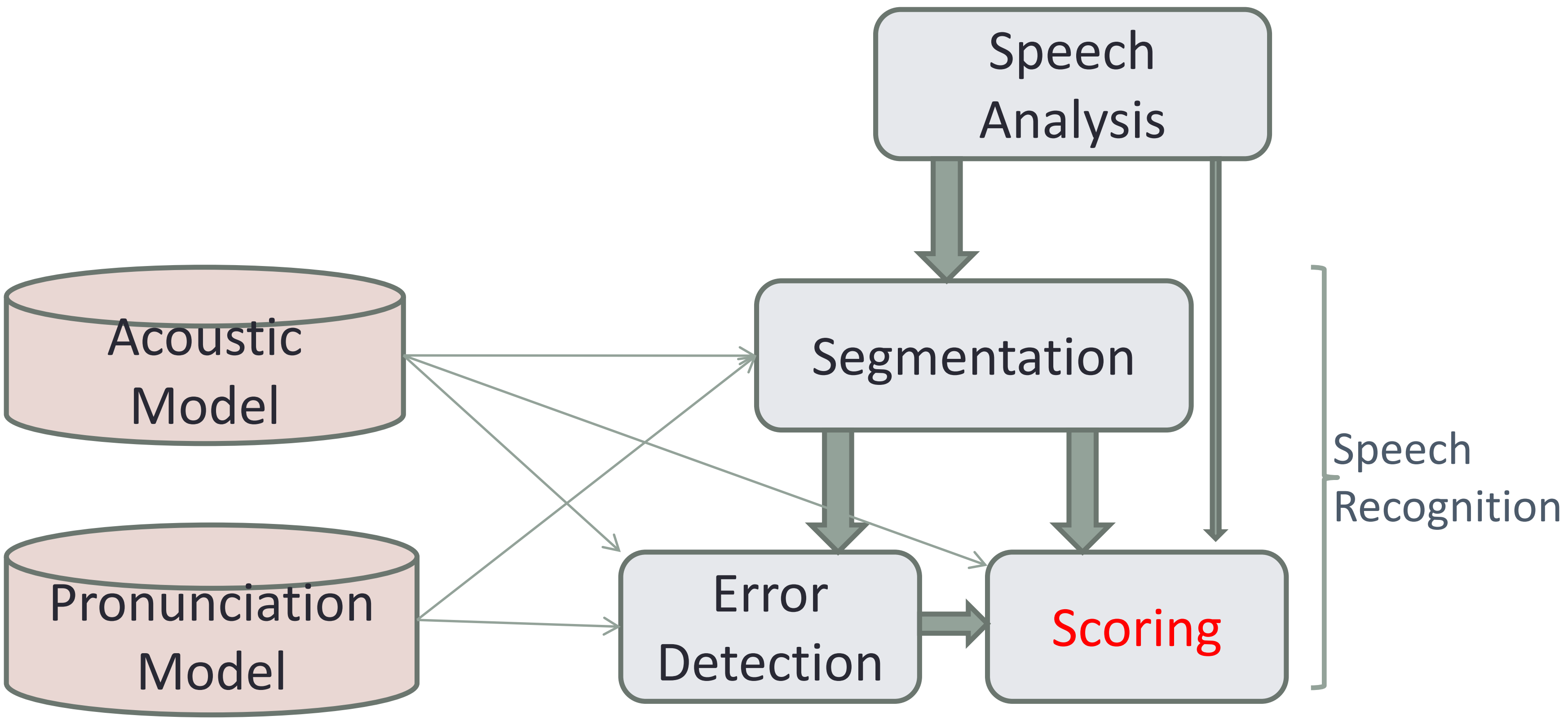
- Not necessarily compute p(x|w'), but test if w' is more likely than w

- Explicit classifier (verifier) learning
  - Incorporate many features
  - Focus on error detection
  - by assuming segmentation
- Linear Discriminant Analysis (LDA)
- Support Vector Machines (SVM)

Other Issues in Error Detection

- Filter and prioritize many (possible) individual phone errors
- error miss >> false alarm
  - Not to discourage learners
- Feedback
  - How to correct errors

Flowchart of Pronunciation Error Detection and Scoring



Scoring: Standpoints

- Native-likeness

  "How close to golden native speakers?"

  → $P(X|W,\lambda_G)$

  - What is the "golden" model? British? American?...
  - Impossible to free from L1 effect, speaker characteristic

- Intelligibility

  "How distinguishable (less confusable) from other phones?"

  → $p(W|X)$

  - Some pronunciation may not be recognized as anything
  - Need to consider L1 phones as well → assume L1

Scoring based on Native-Likeness

- How close to golden native speakers?
  - Defined by $p(X|W,\lambda_G)$ λ_G: golden model
  - Normalized by $p(X|W,\lambda_N)$ λ_N: non-native model
  - In summary, likelihood ratio

$$\frac{p(X|W,\lambda_G)}{p(X|W,\lambda_N)} \approx \boxed{\prod} \frac{p(x|w,\lambda_G)}{p(x|w,\lambda_N)} \approx \prod \boxed{\prod_t} \frac{p(x_t|w,\lambda_G)}{p(x_t|w,\lambda_N)}$$

Mean w.r.t. phones Mean w.r.t. time-frame

Π: geometric mean= arithmetic mean in logarithm

Scoring based on Intelligibility

- How distinguishable (less confusable) from other phones?
  - Measured by p(W|X)

$$\frac{p(X|W)}{\sum_{W'} p(X|W')} \approx \prod \boxed{\frac{p(x|w)}{\sum_{w'} p(x|w')}}$$

posterior prob.

$$\approx \prod_t \frac{\boxed{p(x_t|w)}}{\boxed{\max p(x_t|w')}}$$

forced alignment

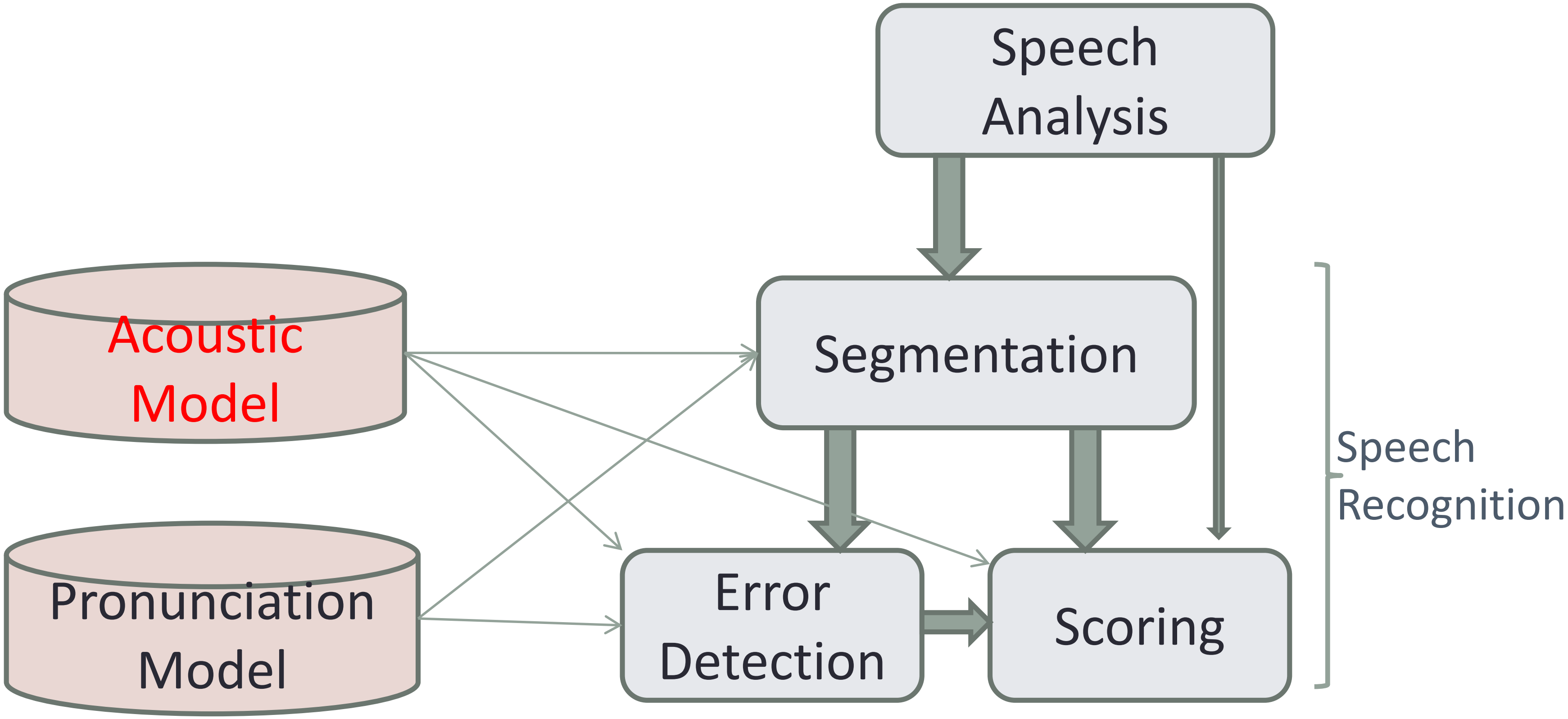
Viterbi score

  - Often called GOP (Goodness Of Pronunciation)
    - becomes 1 if best w'=w
  - Need to adapt to non-native speech
  - Need to consider L1 phones

Scoring to Assessment

- Other factors
  - Duration modeling & evaluation
  - Other prosodic aspects…accent, intonation
  - Speech rate
- Score mapping
  - Linear regression to fit to human rater's evaluation

Flowchart of Pronunciation Error Detection and Scoring



Acoustic Modeling: Native vs. Non-native

- Native speech
  - “Gold standard”, but does not match
- Non-native speech
  - Matched, but error-prone
  - There is not large database available

⇩

- Adaptation from native to non-native
- Phone model of L1 is used for the same phone (in the IPA inventory)

Context-Independent Modeling

- Context-dependent (e.g. triphone) models are widely used in ASR
- Context-independent (monophone) model works well, even better, in CALL
  - Phonetic context is not reliable in non-native speech ← insertion of vowels
  - Better segmentation accuracy even in native speech

Speaker Adaptation of Acoustic Model to Non-native Speech

- Pronunciation of adaptation data may not be correct
- Compare baseform label (automatic but error prone) and hand label (correct but costly)
- Phone accuracy: measured based on hand-label including errors

[Tsubota 2004]

Acoustic model (native model)	Phone accuracy
No adaptation	75.4
Hand label	81.0
Baseform label	80.6

Lexicon baseform label is sufficient

Acoustic Model: Native model vs. Non-native model

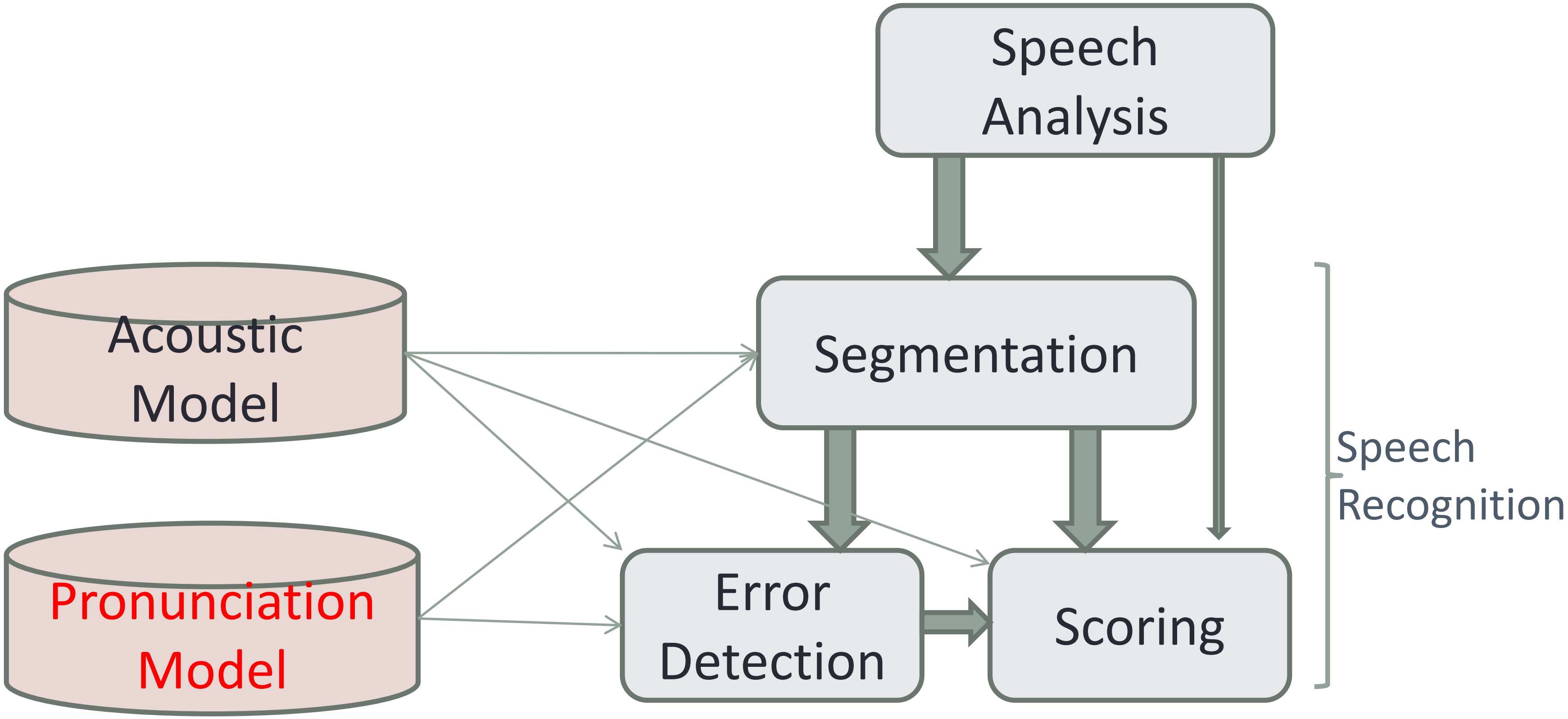
- Non-native speech database (MEXT project)
  - 13129 utterances by 178 speakers
  - Pronunciation errors are not annotated (too costly)
  - Dictionary label vs. automatic label with ASR
    - Both are error prone

[Tsubota 2004]

Acoustic model	baseline	speaker adapt
Native English model	75.4	80.6
Non-native model (baseform)	78.0	81.8
Non-native model (ASR)	77.1	81.5

- Non-native model is more effective, even with dictionary label
- The superiority is reduced with speaker adaptation

Flowchart of Pronunciation Error Detection and Scoring



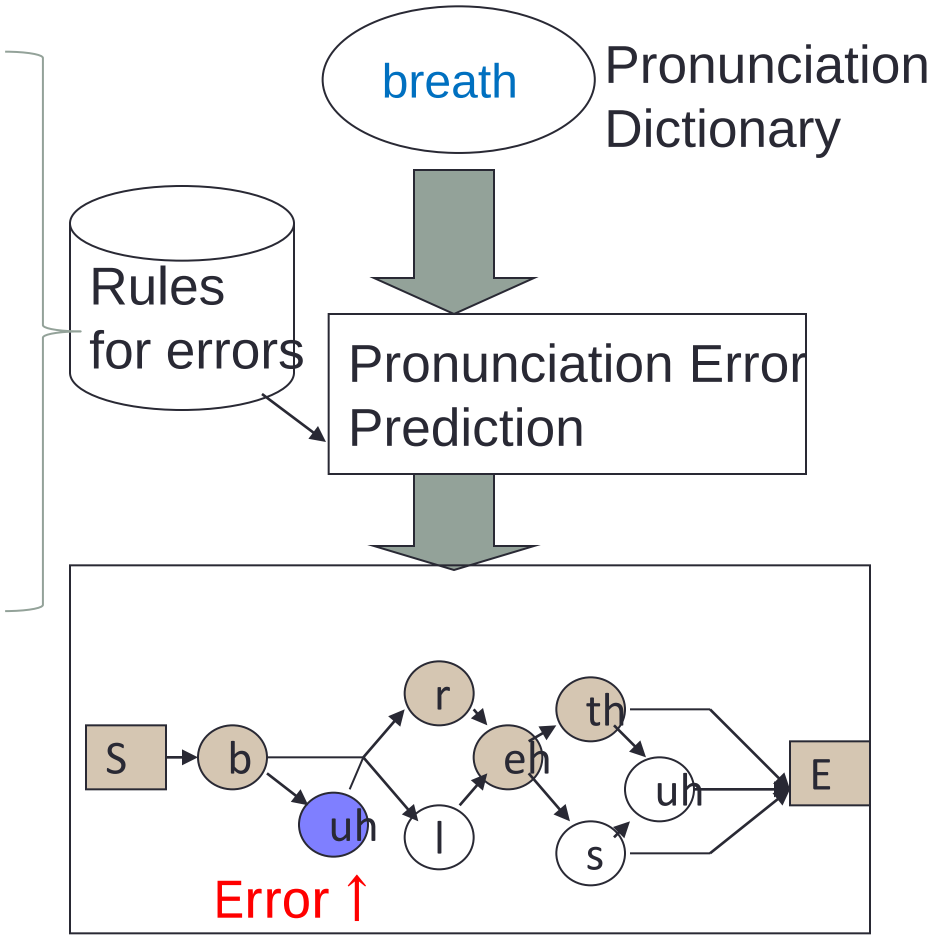
Pronunciation Model

- Standard baseform → possible errors
- Constraint of L1 is effective
- Linguistic knowledge
  - /v/ → /b/, /ϑ/ → /s/
  - Substitution with similar phone of L1
  - Insertion of vowels

- For GOP score computation, simple phone loop model (=no pronunciation model) is used

Error Prediction in Pronunciation Model

- No equivalent syllable in L1
  (ex.) sea → she
- No equivalent phoneme in L1
  (ex.) l → r, v → b
- Vowel insertions
  (ex.) b-r → b-uh-r



Pronunciation Model Training

- Hand-craft phonological rules
  - Expert knowledge needed
  - Too many rules cause false alarms, degrading recognition performance
  - Tradeoff between coverage and perplexity

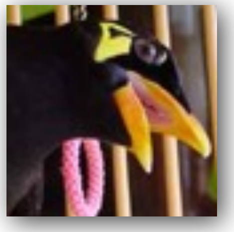
- Machine learning from annotated data
  - Statistical learning of rewriting rules [Meng 2011]
  - Decision tree to find critical rule set [Wang 2009]

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL System (TK)
- Database for CALL (NM)

Another approach for segmental assessment

- Pronunciation training is not impersonation training [Minematsu'07].
  - Impersonation = trying to speak exactly like a target speaker
    - Not needed for pronunciation training.
    - Students are not myna birds!!



- Likelihood scores are impersonation scores, not pron. scores.
  - $P(o|p)$ = similarity bet. a student's ***p*** and the mean speaker's ***p*** in training data.
  - Inadequate if a student is a child and HMMs are trained from adult teachers.
- Posterior probability (GOP) is a score with normalization.

$$P(p|o) = \frac{P(o|p)P(p)}{\sum_q P(o|q)P(q)} \approx \frac{P(o|p)}{\max_q P(o|q)}$$

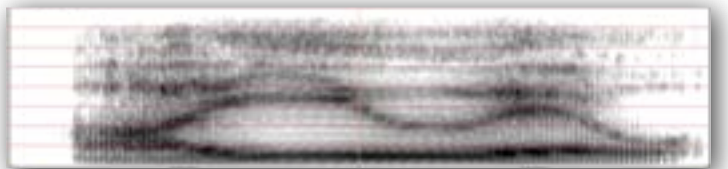
forced alignment

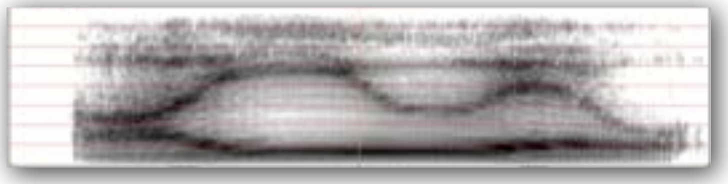
cont. phoneme recognition

  - But alignment and recognition fails due to mismatch bet. students and teachers.
    - Then, speaker-adapted HMMs are often used or native children's data are collected.
    - So, posterior probability is a score of impersonation, again?

Another approach for segmental assessment

- The essential problem lies in the use of spectrum envelopes.
  - SE carries information both of linguistic content and speaker identity.


Tall speaker


Short speaker

- But students imitate only the linguistic content!
  - Speaker information in the teacher's utterance is ignored by students.
    - What does "Hcopy" copy from utterances? What do students copy from utterances?
  - How to make a machine ignore the speaker component in an utterance?

- What is the commonly observed speech pattern?
  - Among linguistically identical but acoustically different utterances.
  - This pattern is the target of students' imitation but what is that?

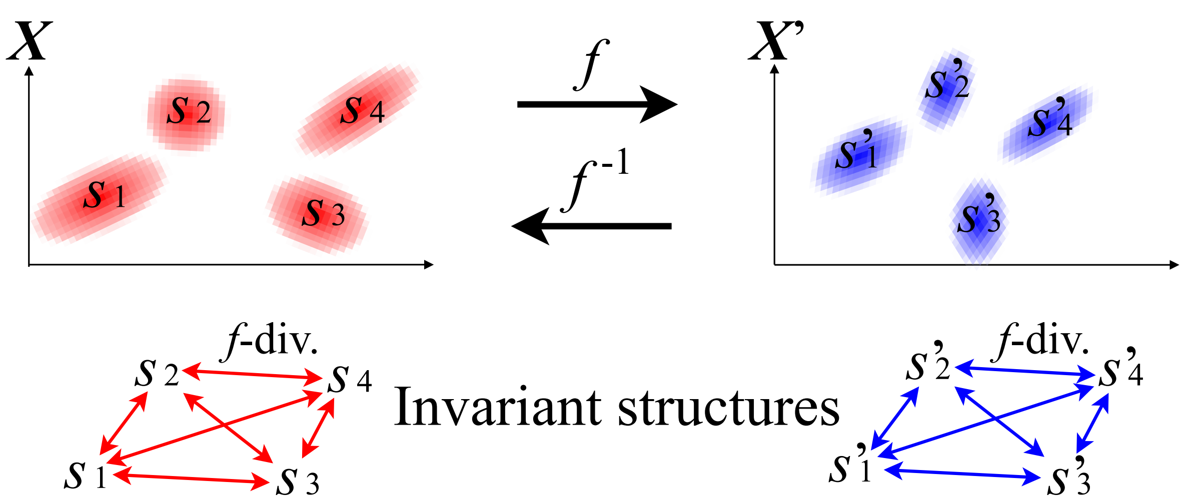
Another approach for segmental assessment

- Speaker difference is often modeled as feature space transformation.
  - The question is “what are transform-invariant patterns or features?”
  - *f*-divergence is invariant with any kind of invertible transform (sufficiency).
    - The invariant features have to be *f*-divergence (necessity). [Qiao+’10]

$$f_{div}(p_1, p_2) = \oint p_2(x) g\left(\frac{p_1(x)}{p_2(x)}\right) dx$$

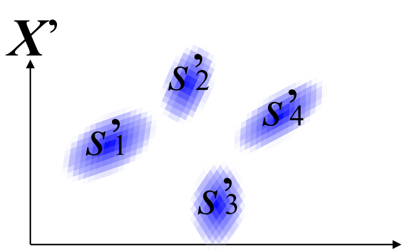
KL-div, Bhattacharyya distance ∈ *f*-div.

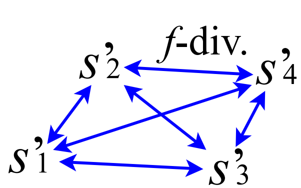
- From individual sounds to their sound system [Minematsu’04]
  - Each sound is dependent on speaker but their system is independent of speaker.
  - Any event has to be characterized as distribution not as point.

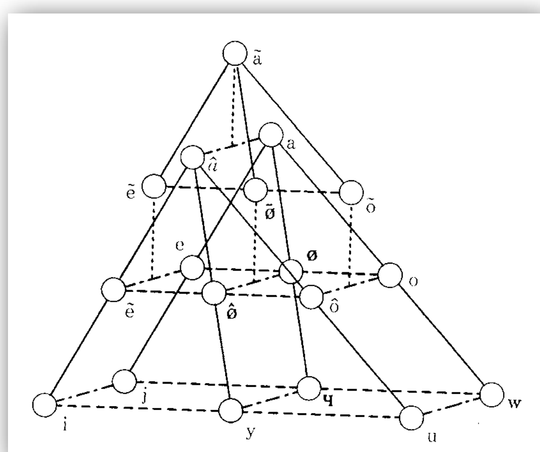


Another approach for segmental assessment

- From individual sounds to their sound system [Minematsu+’06]
  - It should be focused on whether *the native sound system* is found in a student’s utterances not whether *native sounds* are found there.
- From phonetics to (structural) phonology
  - Acoustic phonetics focus on acoustic features of individual phones.
  - Structural phonology focuses on features of their sound system.
  - Roman Jakobson (1896-1982)
    - The sound shape of language (1987)
    - *We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.*







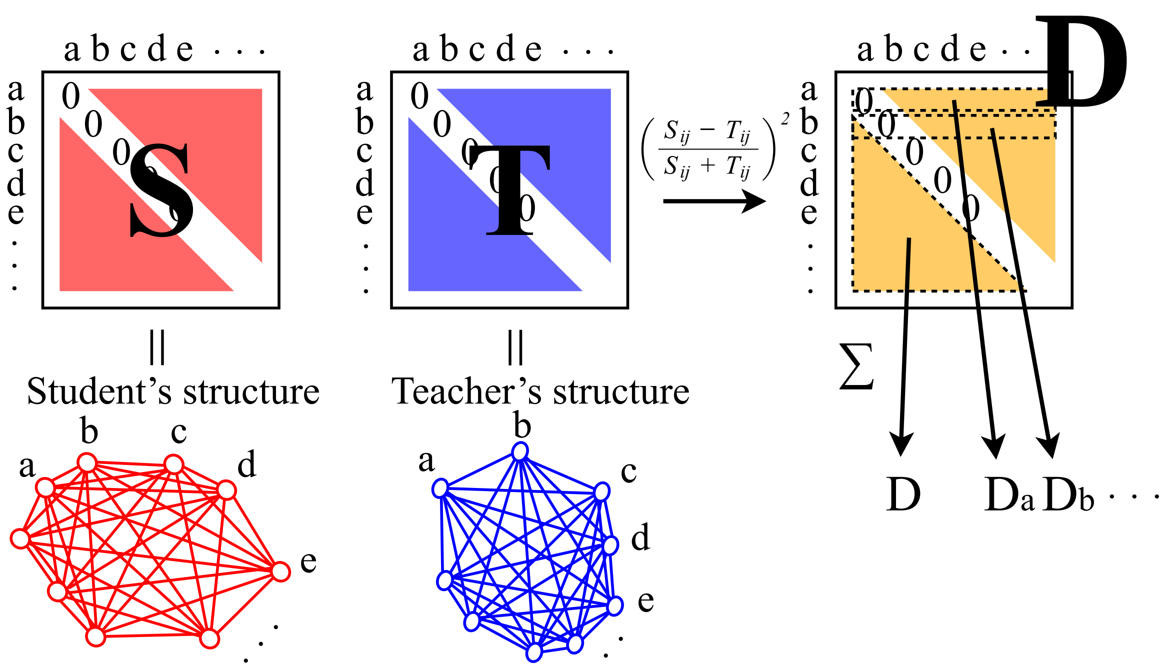
Pronunciation structure

- Topological difference between a student and a teacher
  - Speaker-dependent phoneme HMMs are build.
    - Phoneme-based *f*-div. distance matrix is calculated from a student and a teacher.
    - S : matrix from a student, T : matrix from a teacher
  - $\boldsymbol{S - T = D}$: difference matrix between S and T

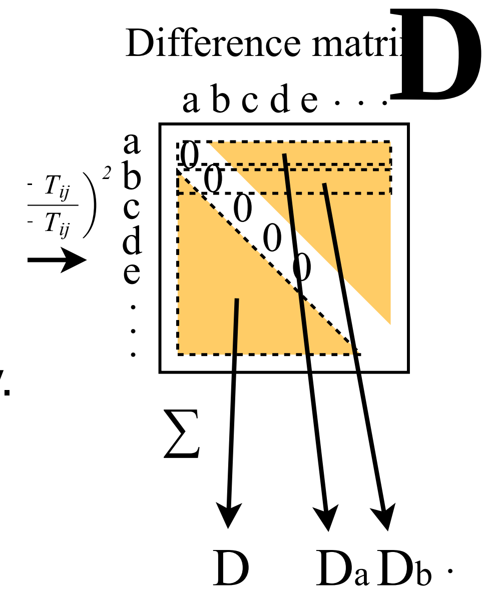
$$\boldsymbol{S_{ij}, T_{ij}} = \sqrt{\text{Bhattacharyya distance bet. two phonemes}}$$ BD ∈ f-div.

$$\boldsymbol{D_{ij} = (S_{ij} - T_{ij})^2}$$ [Minematsu+'06]

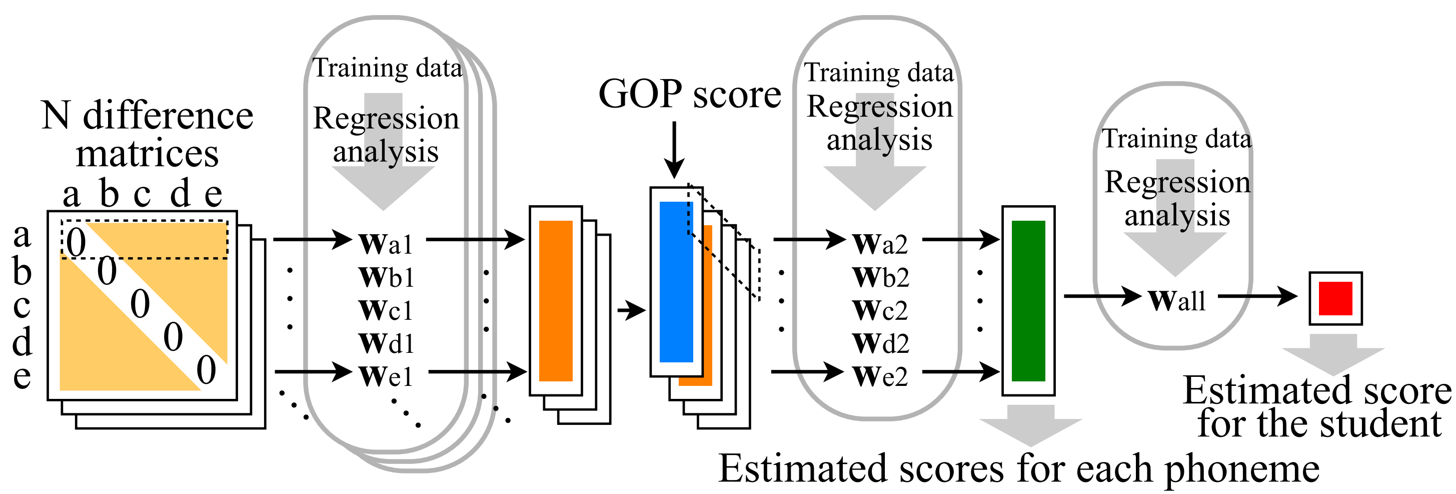
$$\boldsymbol{D_{ij} = \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}}\right)^2}$$ [Suzuki+'10]



Pronunciation structure

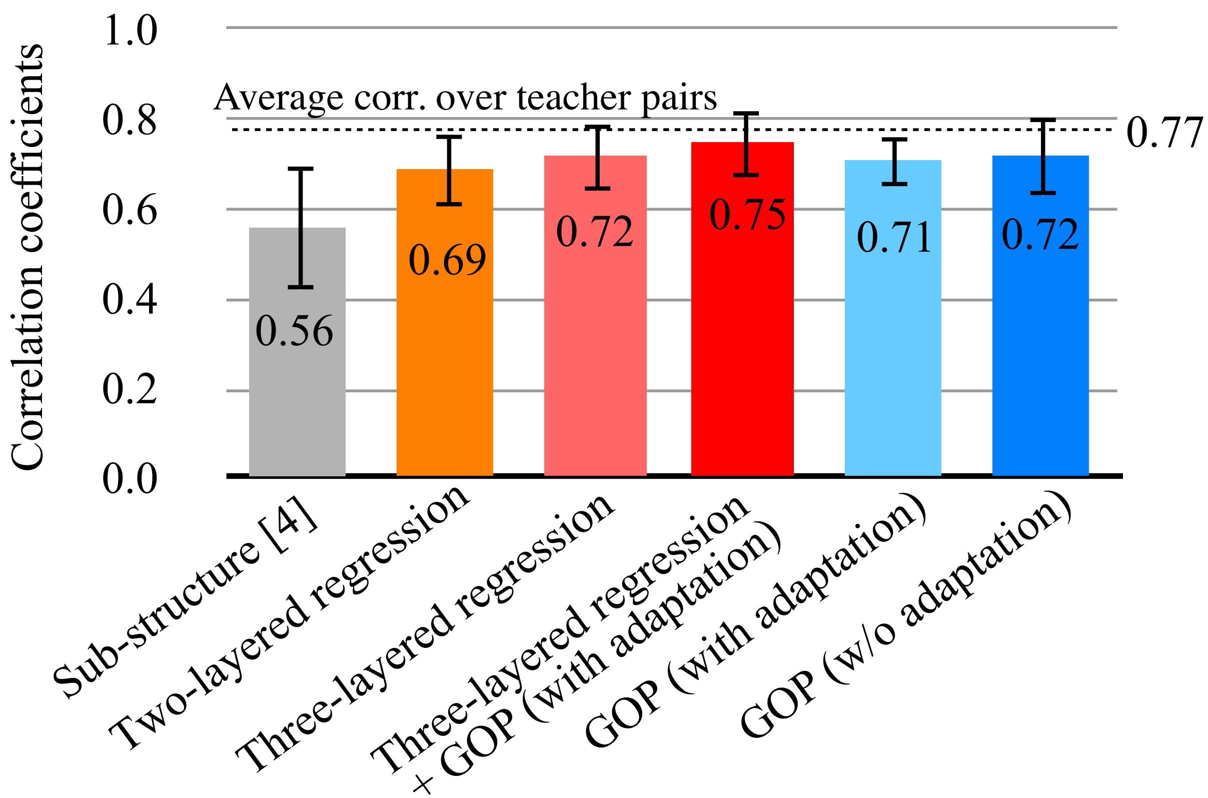


- Global Assessment Score calculated from D matrix
  - $\mathbf{GAS} = \sum_{i<j} D_{ij}$ [Minematsu+'06]
    - Very effective when the target sounds are vowel-like sounds only.
    - Not effective when all the phonemes are considered.
  - $\mathbf{GAS} = \textbf{weighted sum of } D_{ij}$ [Suzuki+'10]
    - Can treat all kinds of phonemes well.
    - Not simple linear regression but multilayer linear regression is applied.
    - D matrices obtained from different teachers (features) can be used additionally.
    - Phoneme-based GOP scores can be used additionally.



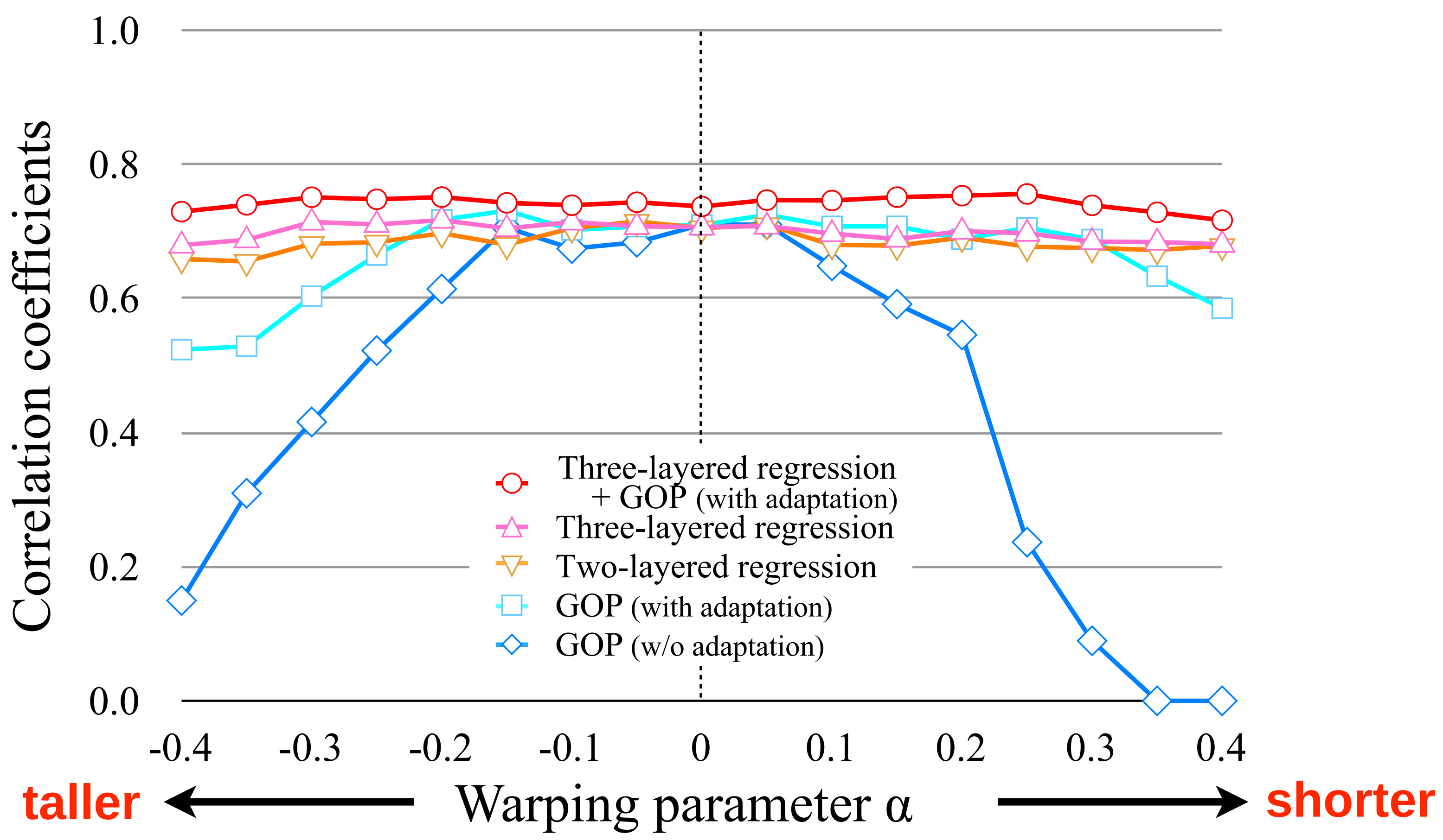
Pronunciation structure

- Experiment using pronunciation structures [Suzuki+'10]
  - About 60 utterances per student (teacher) to train a spk-dependent HMM set.
  - Number of teachers used for the experiment
    - Two-layered regression : only 1 male teacher
    - Three-layers regression : only 1 male and 1 female teachers
  - Correlation between human teachers' scores and machine scores



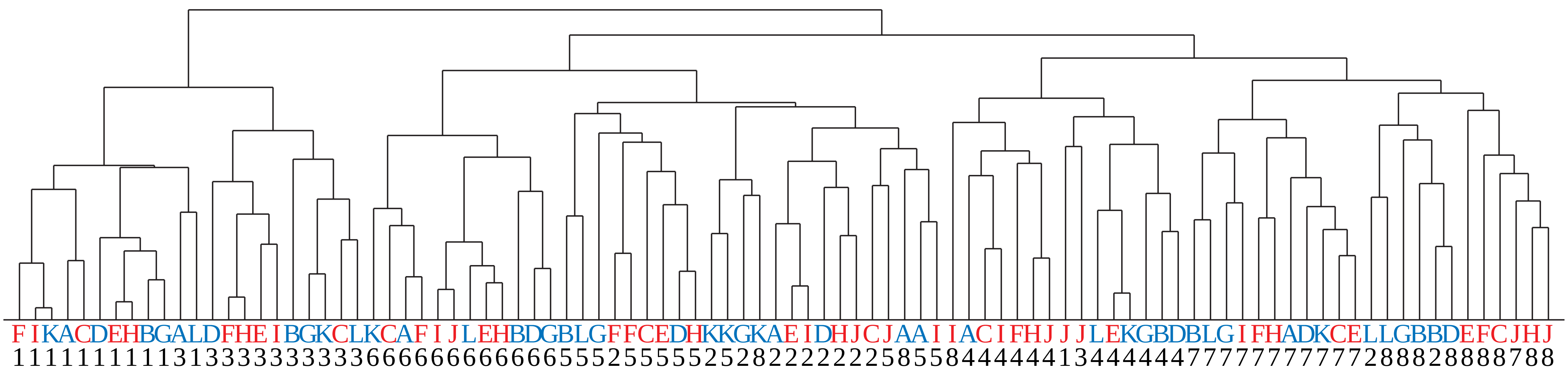
Pronunciation structure

- Experiment using warped utterances [Suzuki+'10]
  - Simulated very tall students and very short students.
  - Only a single teacher is used in the two-layered regression.



Learner clustering based on their pron.

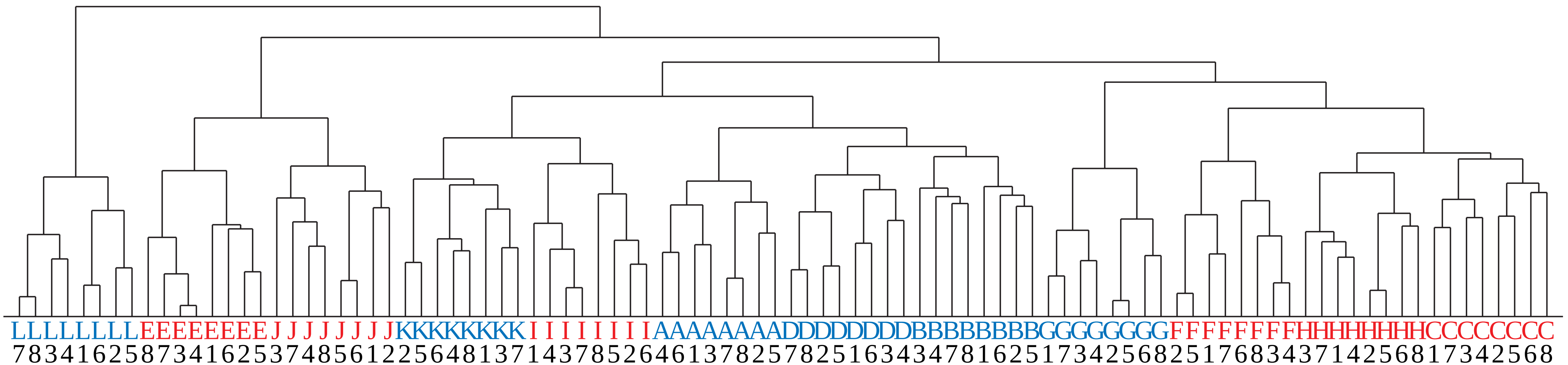
- Clustering “simulated” 96 students [Minematsu+’06,’07]
  - Only vowels are focused.
  - Speakers are 12 very good learners of American English (spk-A to spk-L).
  - They are asked to produce AE vowels and JE vowels, uttered in word context.
    - 7 differently accented vowel structures and a good and normal vowel structure.
    - 1-7 : Japanese accented structures, 8 : non-accented structure
    - ex) ɑ, ʌ, æ, ə, ɚ, ɔ, ɛ, ɪ, i, ʊ, u (Red vowels are replaced by Japanized versions.)
  - 12 students x 8 pronunciations = 96 simulated students



A to L : student ID,   1 to 8 : pronunciation ID

Learner clustering based on their pron.

- Acoustic clustering vs. structural clustering [Minematsu+'06,'07]



A to L : student ID,   1 to 8 : pronunciation ID

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL System (TK)
- Database for CALL (NM)

Basic prosodic features

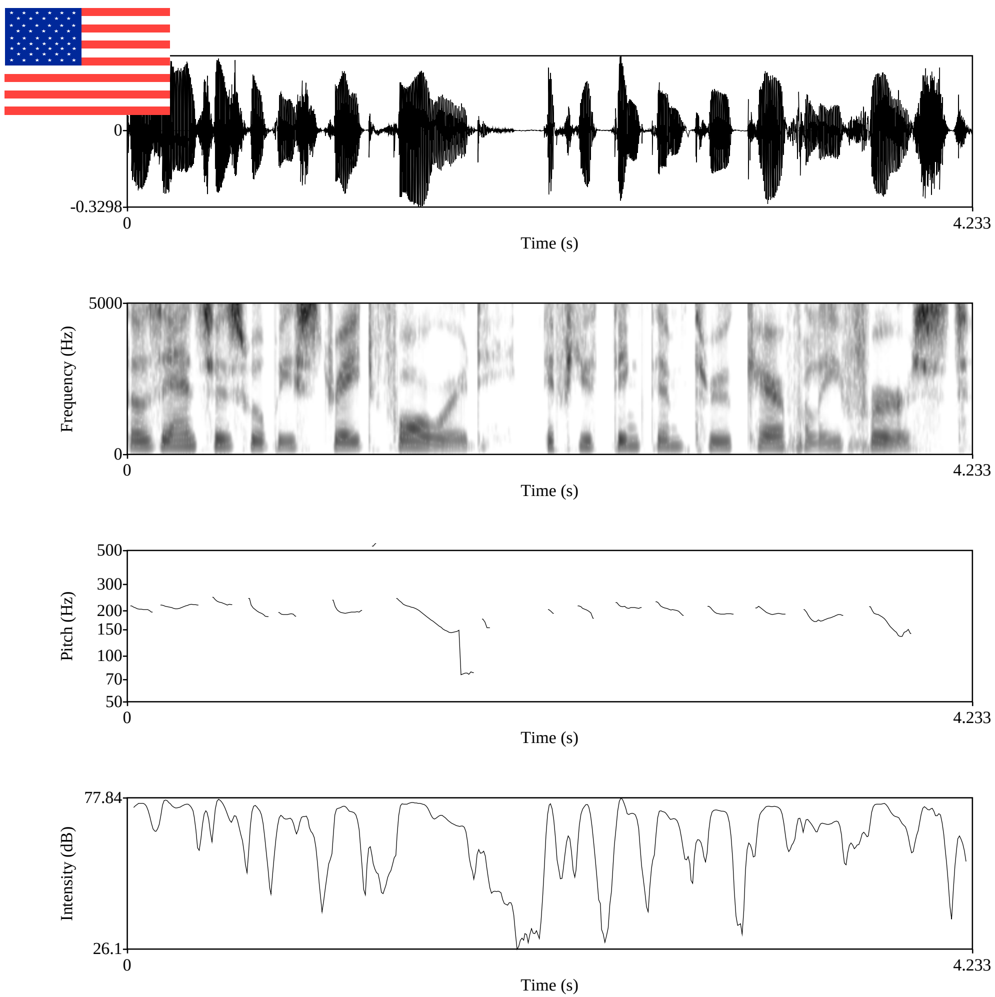
- Three basic psychological terms and their acoustic correlates

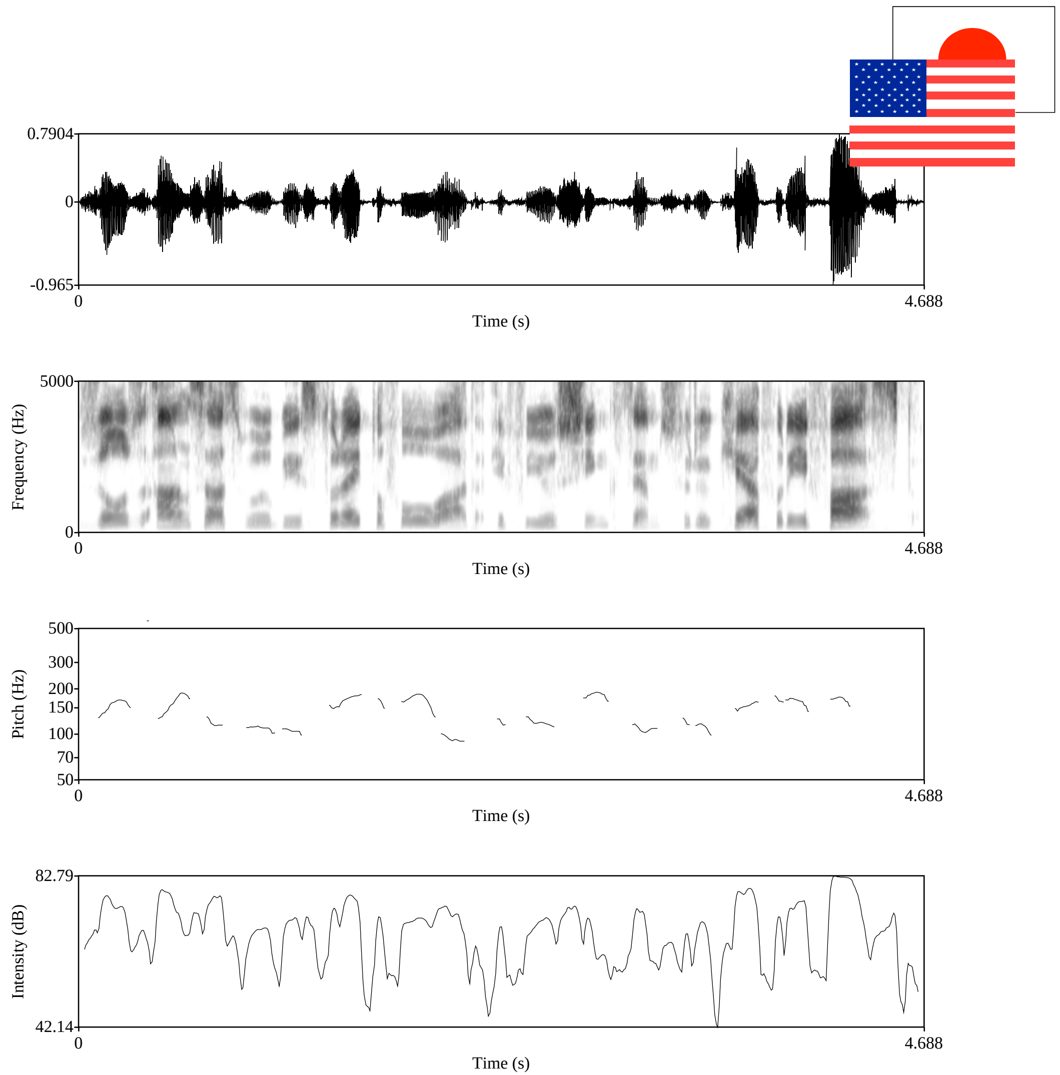
psychological	physical (acoustic)	related phenomena
pitch	fundamental frequency	intonation, word accent speaker identity
loudness	energy, intensity, sound pressure level	word accent (word stress)
duration*	duration*	rhythm
timbre	spectrum envelope	phoneme speaker identity

  - It seems that two distinct terms are not prepared well for perceptual length and physical length of a sound.
- Foreign accent and prosodic features
  - Various types of prosodic deviation can be found in foreign accented speech depending on the native language of a learner and the target language.

Basic prosodic features

- "Those answers will be straightforward if you think them through..."
  - Results of acoustic analysis using Praat.
    - http://www.fon.hum.uva.nl/praat/





Prosodic assessment of pronunciation

- Use of various prosodic metrics to estimate prosodic quality
  - Duration-based metrics to predict “fluency” [Cucchiarini+’98,’02]
  - Model-based and non-model based prosodic metrics [Maier+’09][Huang+’10]
- Additional prosodic features used to estimate overall proficiency
  - Duration log-likelihood [Kim+’97], rate of speech [Franco+’00]
  - Linear combination of various scores to predict proficiency [Hirabayashi+’10]
- Word accent (word stress) generation assessment
  - Position [Minematsu+’97][Imoto+’99] and manner [Minematsu+’00]
- Rhythm assessment
  - Rhythm metrics [Ramus+’99,’02][Grabe+’99,’02]
- Intonation(+energy) pattern comparison bet. a student and a model
  - Word-based comparison [Suzuki+’08][Cheng+’11]
  - Multiple units for comparison [Yamashita+’05]
- Corrective feedback generation
  - Decision-tree based generation [Liao+’10], using a learner’s voice [Hirose+’03]

Duration-based metrics

- Duration-based metrics to predict "fluency" [Cucchiarini+'98,'02]
  - 60 non-native learners of Dutch and 20 native speakers
  - Forced alignment using an ASR engine
  - 3 groups of raters, 3 raters per group (phonetician, therapist1, therapist2)
  - Fluency assessment was done for each material (sentence?).

	intrarater reliability			**interrater reliability**
	rater 1	**rater 2**	**rater 3**	
ph	.97	.94	.95	.96
st1	.94	.97	.96	.93
st2	.90	.76	.91	.90

Table 1 Intrarater and interrater reliability coefficients (Cronbach's alpha) for the three rater groups, ph, st1, and st2.

Duration-based metrics

- Duration-based metrics to predict “fluency” [Cucchiarini+’98,’02]
  - Each duration-based metic is examined as its ability to predict “fluency”.

- ros = rate of speech: # segments / total duration of speech plus sentence-internal pauses
- ptr = phonation/time ratio: total duration of speech without pauses / total duration of speech plus sentence-internal pauses
- art = articulation rate : # segments / total duration of speech without pauses
- tdp = total duration of sentence-internal pauses: all silences longer than or equal to 0.2 sec
- alp = average length of pauses
- #p = # of silent pauses
- mlr = mean length of runs: average number of phones occurring between unfilled pauses of not less than 0.20 secs
- #fp = # filled pauses: ə, əm
- #dy = # dysfluencies (repetitions, restarts, repairs)

	Phoneticians	**Speech therapists 1**	**Speech therapists 2**
ros	.93	.91	.90
ptr	.86	.88	.89
art	.88	.84	.81
#p	-.84	-89	-.89
tdp	-.81	-.86	-.86
alp	-.65	-.62	-.65
mlr	.85	.86	.88
#fp	.34	.33	.38
#dy	.42	.48	.40

Table 5 Correlations (corrected for attenuation) between the fluency ratings by the three rater groups and the quantitative measures.

Speech rhythm metrics

- The three rhythm classes
  - Stress-timed languages: English, German, Dutch, Portuguese, etc
  - Syllable-timed languages: French, Italian, Spanish, Cantonese Chinese, etc
  - Mora-timed languages: Japanese, etc
  - X-timed = the *perceptual* interval between two consecutive Xes is constant
    - Stress isochrony, syllable isochrony, and mora isochrony
- Pairwise Variability Index (PVI) [Grabe+'99,'02]
  - Raw PVI (rPVI) and normalized PVI (nPVI)

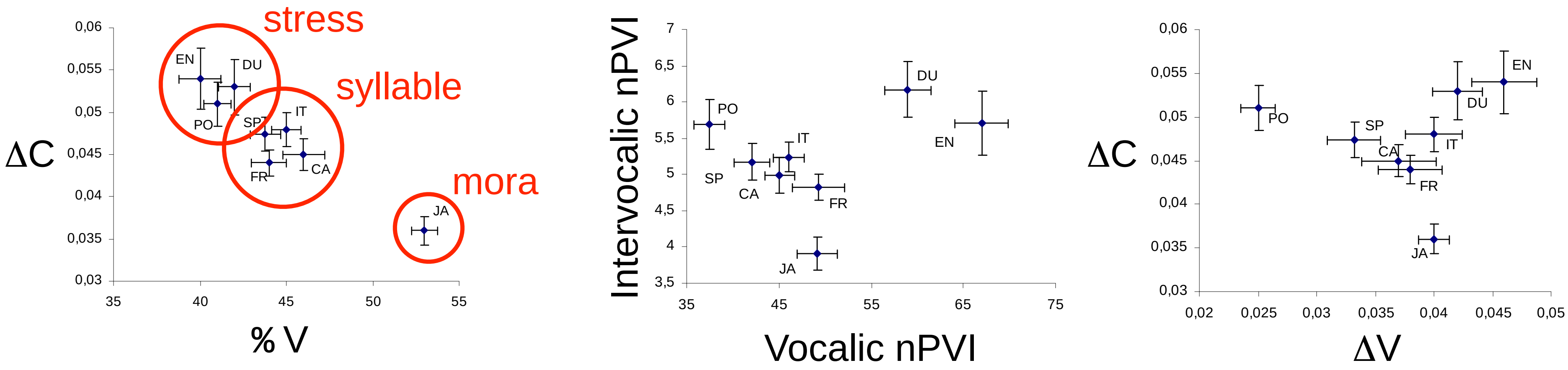
$$rPVI = \frac{100}{m-1} \sum_{k=1}^{m-1} |d_k - d_{k+1}|$$

$$nPVI = \frac{100}{m-1} \sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2}$$

  - d_k is the duration of the k-th interval. m is the number of intervals.
  - "interval" is the vocalic interval or the consonantal interval.
  - Used to classify input utterances as one of the three rhythm groups

Speech rhythm metrics

- Combination of durational statistics of ΔV, ΔC and %V [Ramus'99,'02]
  - ΔX : standard deviation of the duration of Vowel intervals or Consonant intervals within a sentence
    - X interval: interval of a X or a sequence of consecutive Xes
    - Intervocalic interval: interval of a consonant or a consonant sequence
  - %V : percentage of duration taken up by vowel intervals within a sentence
  - Used to cluster various languages in terms of their rhythmic structure.



Various prosodic metrics

- Lang-independent feature set for prosody evaluation [Maier+'09]
  - Word-based 21 metrics + sentence-based 16 metrics
    - Related to F0, energy, and duration
    - 37 metrics x [max, min, mean, std] = 148 features
  - Text-independent 187 prosodic features
  - Support vector regression to predict prosodic quality

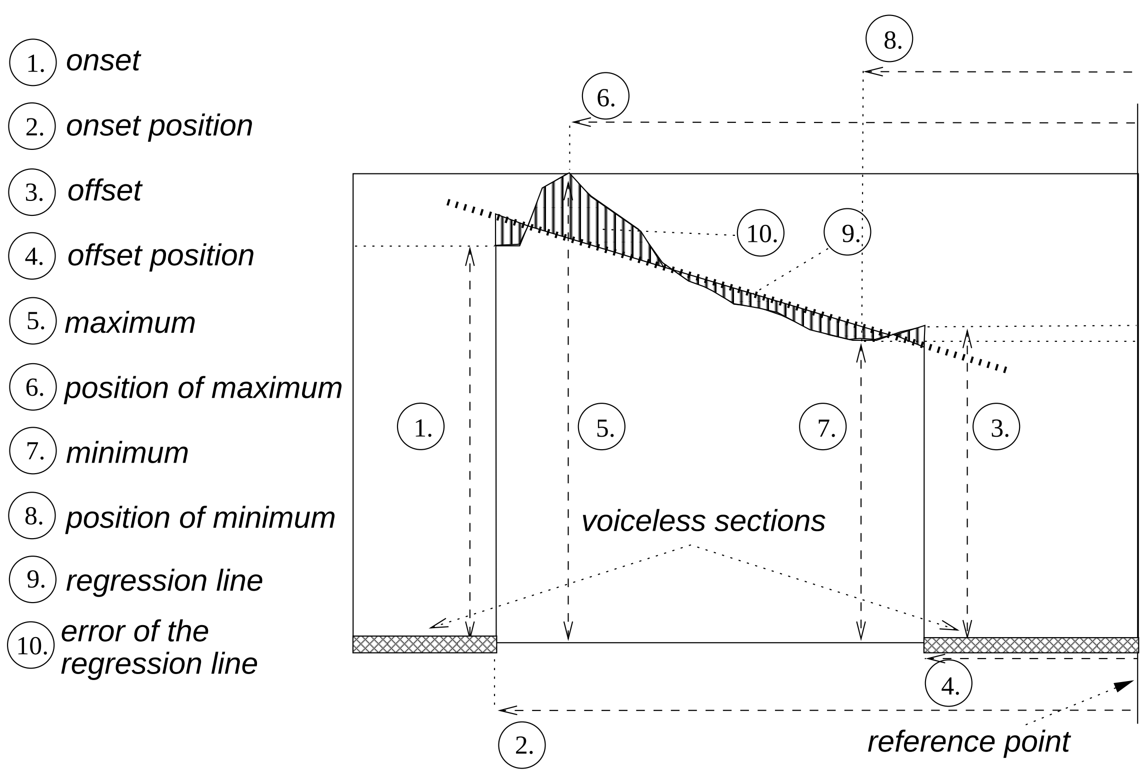


Table 1: *Correlations between the automatic evaluation system and the human raters in comparison to the inter-rater correlation*

language	inter-rater	word-based SVR	word-based SVR (CFS)	text-independent SVR	text-independent SVR (CFS)
German	**0.88**	0.89	**0.92**	**0.88**	0.75
Japanese	**0.92**	-	-	0.76	**0.83**

Linear combination of many features

- Prediction of overall naturalness [Hirabayashi+'10]
  - Many possible features are linearly combined to predict pron. scores
    - LL with native HMMs = LL_{native}
    - LL with HMMs adapted into non-native = $LL_{non\text{-}native}$
    - LL obtained with phone-loop grammar and native HMMs = $LL_{best\text{-}native}$
    - LL ratio = LR = LL_{native} - $LL2_{non\text{-}native}$
    - Posterior probability = LR' = LL_{native} - $LL_{best\text{-}native}$
    - Another LL ratio = LR_{adapt} = $LL_{best\text{-}native}$ - $LL_{best\text{-}non\text{-}native}$
    - Another LL ratio = LR_{mother} = $LL_{best\text{-}native}$ - $LL_{best\text{-}mother\text{-}tongue}$
    - Phoneme recognition rates (rates of correct, substitution and deletion)
    - Word recognition results (rates of correct, substitution and deletion)
    - Standard deviation of power and F0
    - Phoneme-based rate of speech

Linear combination of many features

- Prediction of overall naturalness [Hirabayashi+'10]
  - Results of linear prediction of pronunciation scores

Table 2: Correlation between acoustic measures and pronunciation score ("*" denotes a text-independent measure)

Measure	1 sentence	5 sentences	10 sentences
LL_{native}	-0.466	-0.625	-0.669
$\mathbf{LL_{non-native}}$	**-0.638**	**-0.771**	**-0.804**
LR	**0.800**	**0.859**	**0.880**
* LL_{best}	-0.473	-0.613	-0.660
* $\mathbf{LR_{mother}}$	**0.719**	**0.804**	**0.811**
* $\mathbf{LR_{adap}}$	**0.772**	**0.827**	**0.822**
LR'	0.214	0.273	0.349
Phoneme recog($Sub.$)	-0.298	-0.567	-0.662
Phoneme recog($Del.$)	0.056	0.116	0.220
Phoneme recog($Cor.$)	0.299	0.461	0.483
Word recog(WSJ, $Cor.$)	0.102	0.163	0.261
Word recog(EURO, $Cor.$)	0.113	0.256	0.281
* $Power$	-0.066	-0.057	-0.002
* $Pitch(F_0)$	0.495	0.638	0.691
Rate of speech	**0.523**	**0.692**	**0.773**

Table 3: Correlation between combination of acoustic measures and learner's pronunciation score by human raters

Number of sentences for evaluation	1 sentence		5 sentences		10 sentences	
Acoustic measures	CLOSED	SP.OPEN	CLOSED	SP.OPEN	CLOSED	SP.OPEN
$LL_{non-native}$, LR, LR_{mother}, $Power$, Phoneme recog($Del.$)	0.851	0.804	0.910	0.851	0.927	0.864
Word recog(EURO, $Cor.$), LR, $Power$, Word recog(WSJ, $Cor.$)	0.815	0.770	0.902	0.866	0.929	0.884
Word recog(EURO, $Cor.$), LR, $Power$	0.814	0.771	0.893	0.858	0.918	0.887
LL_{best}, LR_{mother}, $Power$	0.819	0.779	0.891	0.853	0.912	0.878

Word stress detection

- Modeling of (un)stressed syllables [Minematsu+'97][Imoto+'02]
  - HMM-based modeling of syllables (C..CVC..C)
    - Syllable structure dependent (V, C..CV, VC..C, and C..CVC..C)
    - Vowel type dependent (short vowels, long vowels, and diphthongs)
    - Vowel position dependent (head, tail and other in a word)

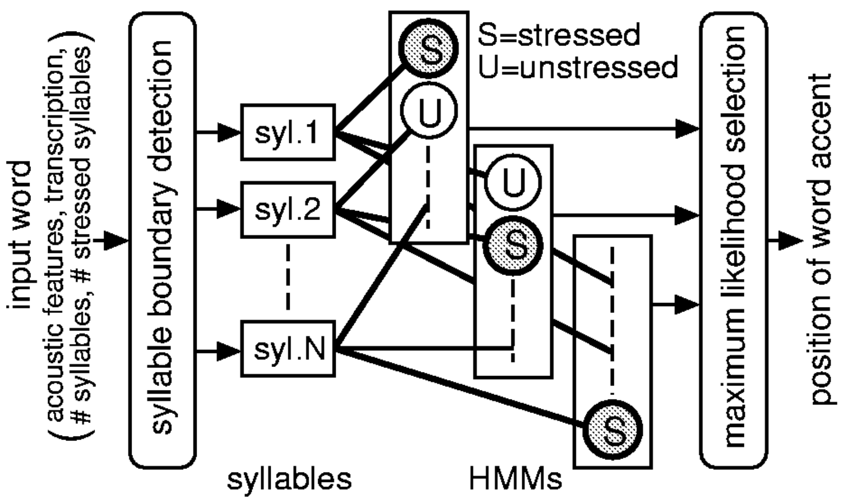


Figure 1: Accent detection using syllable boundaries

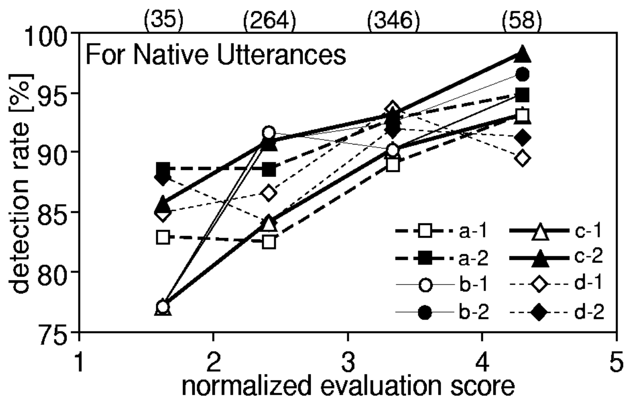


Figure 2: Detection rates for native speakers

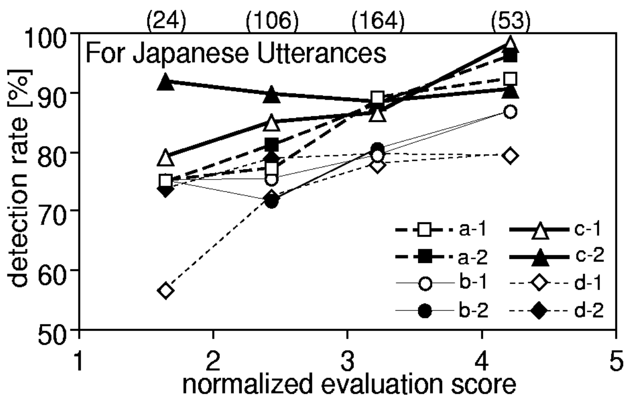
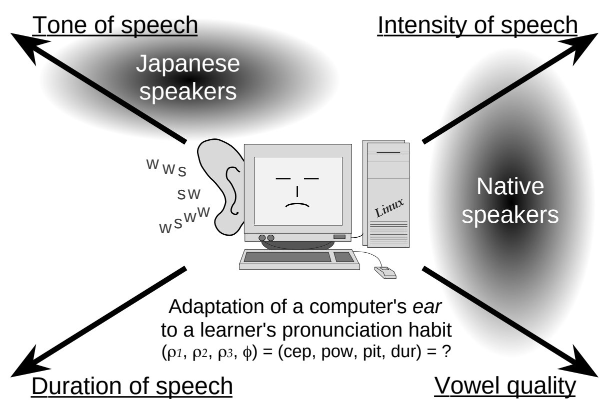


Figure 3: Detection rates for Japanese students

Manner of word stress generation

- Estimation of pron. habit in word stress generation [Minematsu+'00]
  - Word accent in Japanese : pitch accent
    - Fundamental frequency (F0)
  - Word accent in English : stress accent
    - Four multiple factors of F0, duration, power, and vowel quality
    - Japanese tend to produce English word stress mainly by pitch change [Shibuya'96].
  - Stress / unstress identification using multiple weights
    - $P(o|M) = P(F_0|M)^{w_{F_0}} P(dur|M)^{w_d} P(pow|M)^{w_p} P(env|M)^{w_e}$
    - The optimal weights represent the pronunciation habit of individual students.
    - Larger w_{F_0} is observed in word stress generation by Japanese?



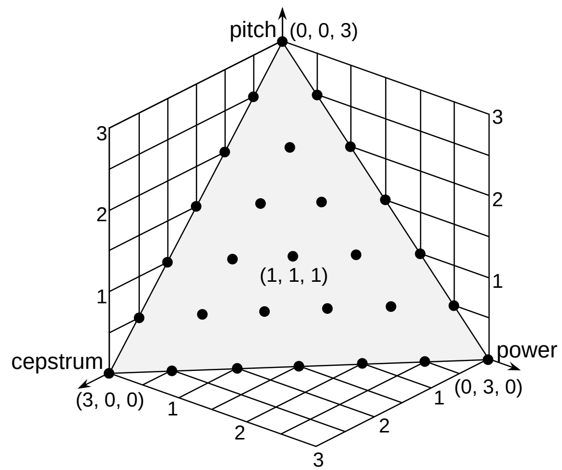
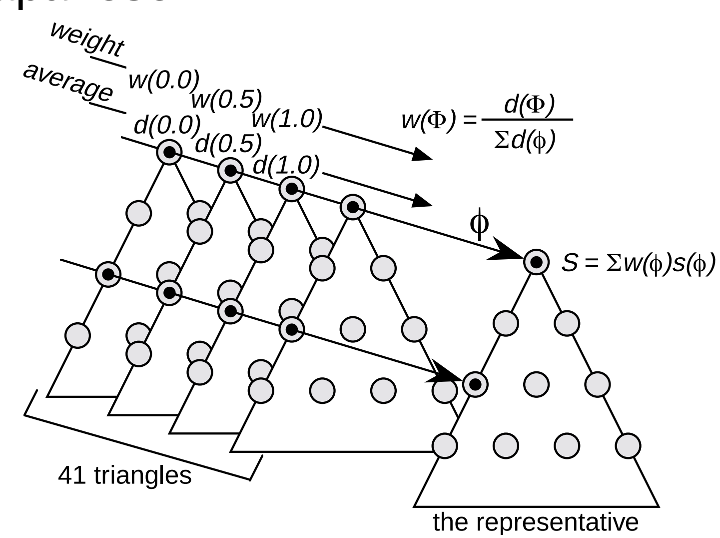
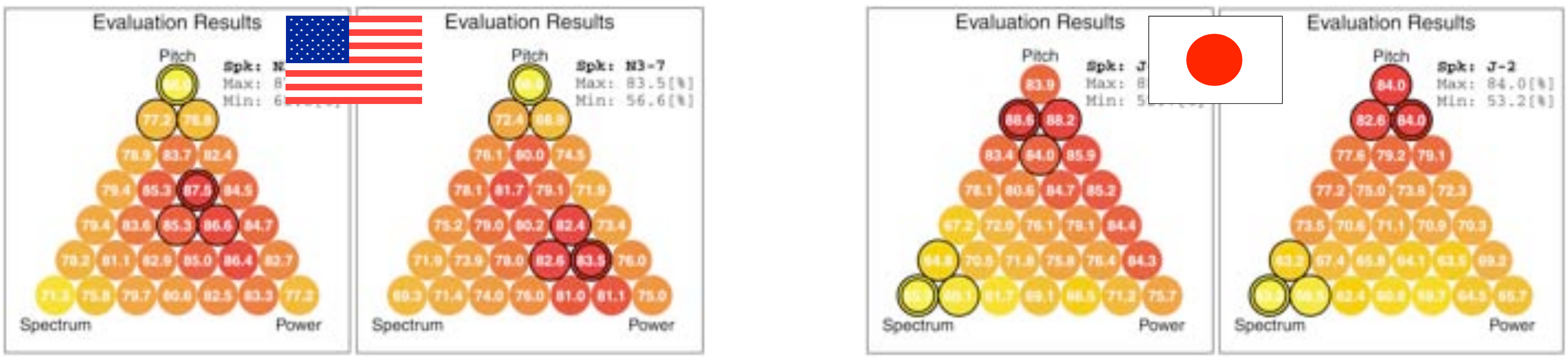


Figure 3: Distribution of the weight combination of $\{\rho_s\}$

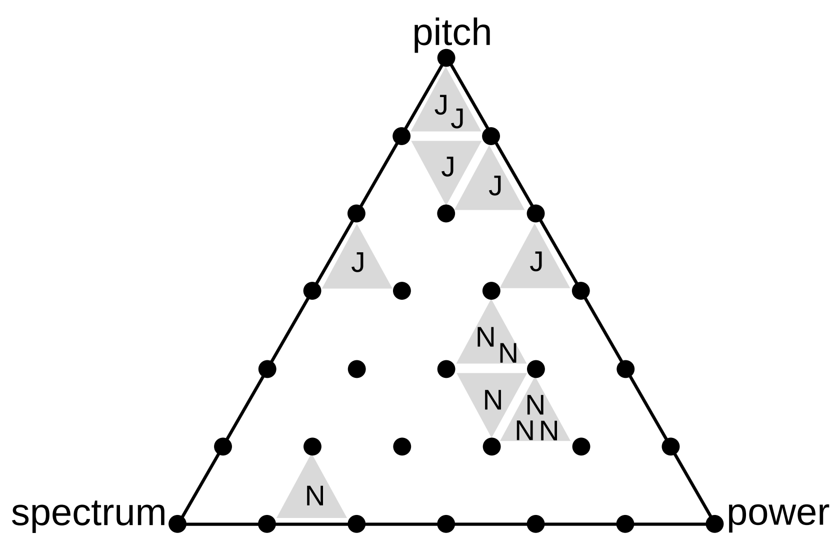


Manner of word stress generation

- Results of pronunciation habit estimation [Minematsu+’00]
  - Four examples of estimation results: two natives and two Japanese



  - Locations of the optimal weights of 7 natives and 6 Japanese students



Tone error detection using DT

- Use of a decision tree to detect tone errors [Liao+'10]
  - A decision tree tells us "why and how the input tone pattern is bad".
    - This info. can be used as easy-to-understand feedback to students.
  - 1 to 5 point human rating scores are converted into binary rating
    - 1/2 = bad and 3/4/5 = good, which are used as labels for training decision trees.
  - A syllable is divided into three segments and F0 mean is obtained from each.
    - F0 differences bet. any segment pair are added to a feature vector.

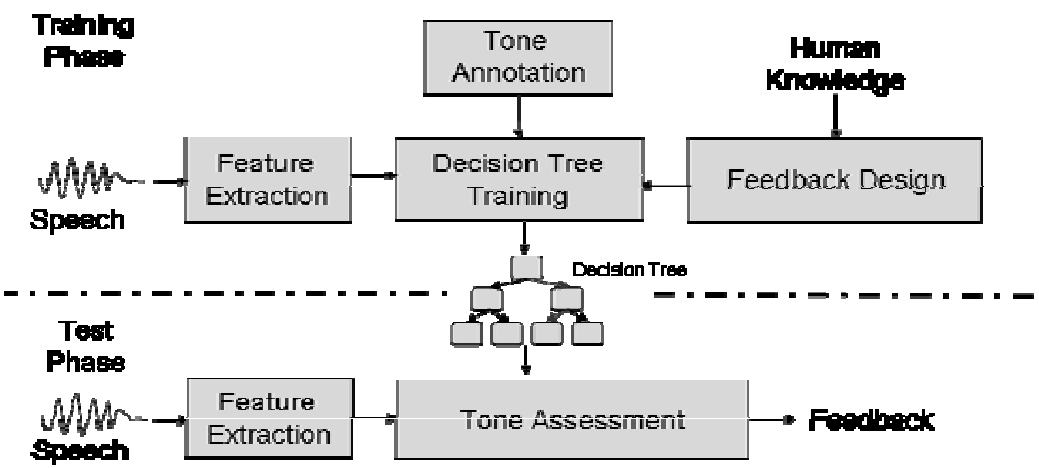


Figure 1: *A decision tree based tone assessment system block diagram with training (upper) and testing (lower) subsystems.*

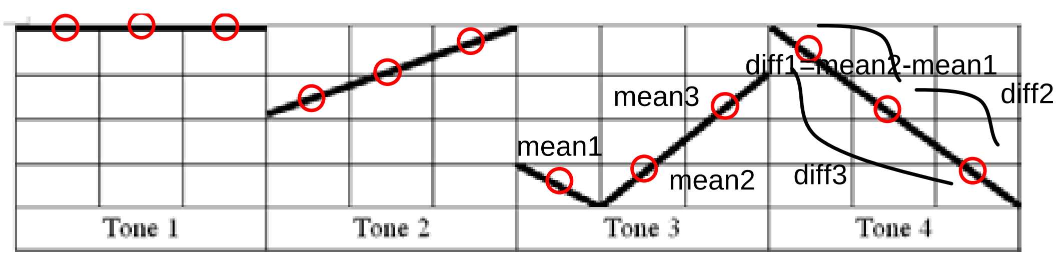
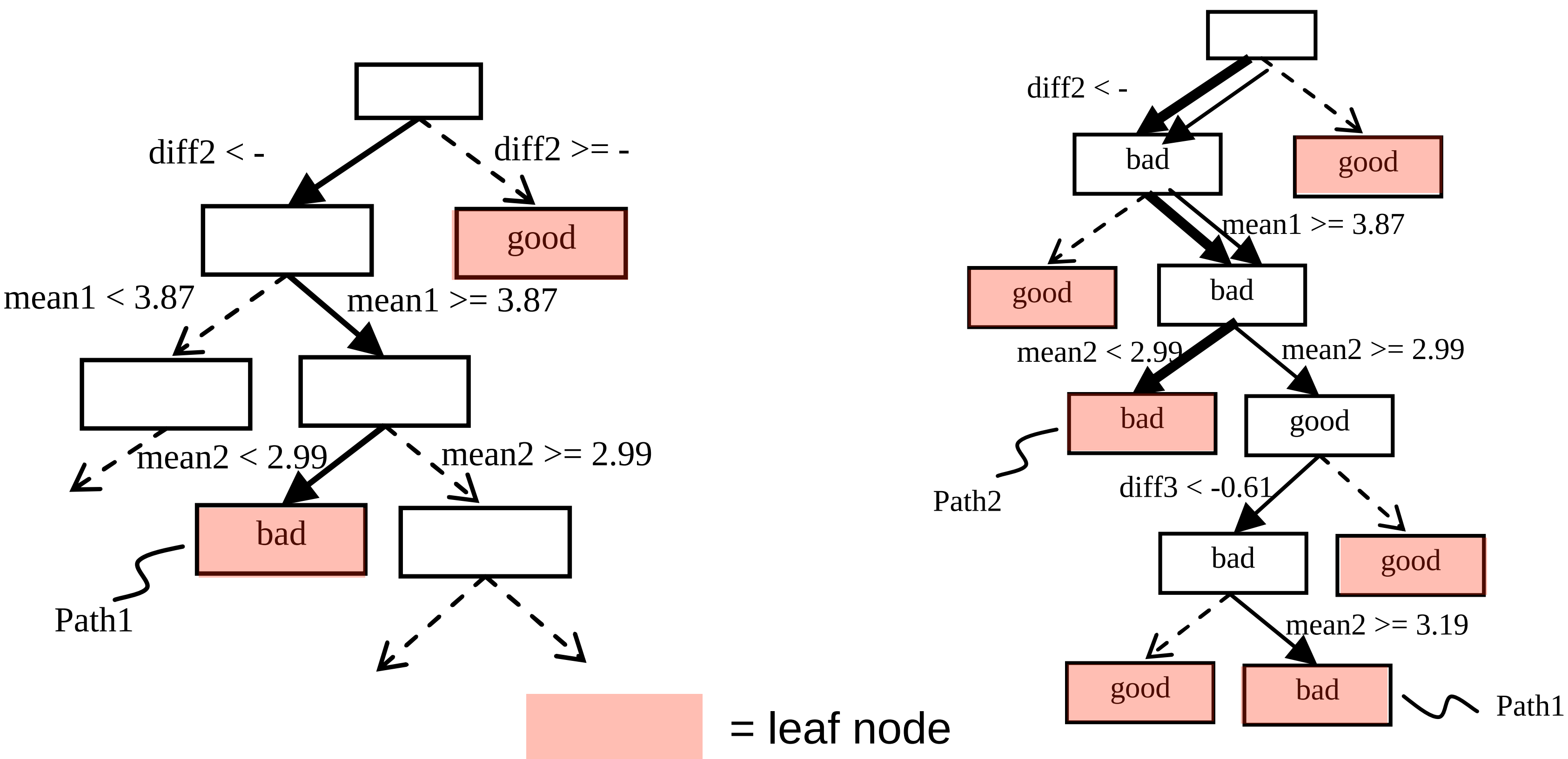


Figure 3: *Illustration of feature vector extraction.*

Tone error detection using DT

- Use of a decision tree to detect tone errors [Liao+'10]
  - Right-context-dependent models are adopted.
  - A set of questions prepared in terms of F0mean and F0diff.
  - Approx. 90% of correct binary judgment (good or bad) for testing data.
  - Potential use of traversed paths for feedback generation



Utterance-based prosodic comparison

- Consideration of characteristics of Japanese English
  - Word-by-word pronunciation [Sugito'98]
  - Too many or too few peak-and-valleys in intonation [Shimizu'95]
- Prosodic comparison between utterances [Yamashita+'05]
  - Multiple units such as word, word boundary, prosodic phrase, and sentence
  - Each unit is determined by phoneme labels obtained from an HMM aligner.

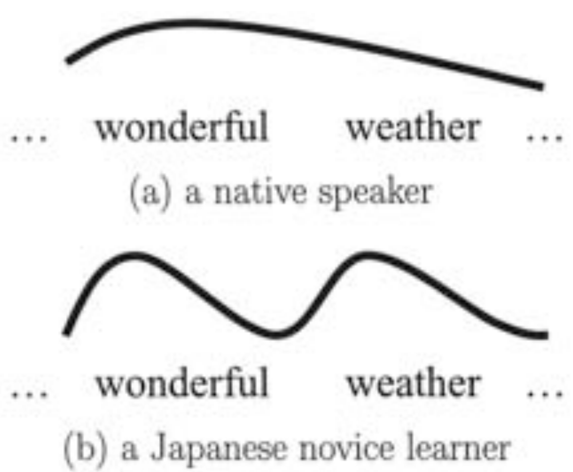
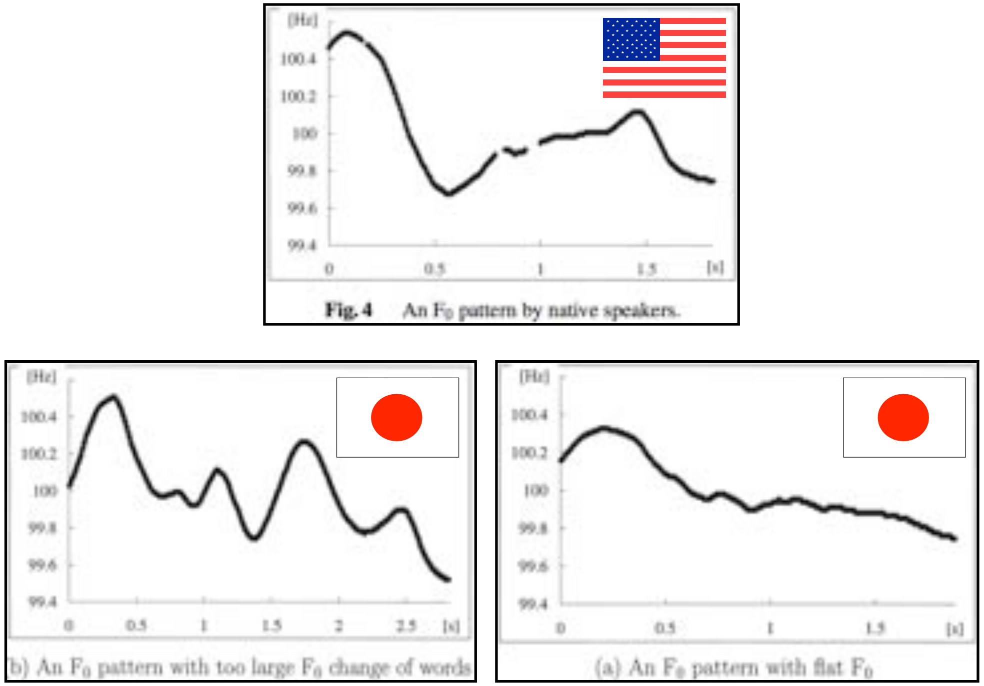


Fig. 1 Typical F_0 patterns at a word boundary.



Utterance-based prosodic comparison

- Prosodic comparison between utterances [Yamashita+’05]
  - F0 contour, power contour, total duration, word duration, pause duration
  - Deviation of an observed contour from its 1-st or 2-nd order approximation.
    - Very low deviation expects that the contour is flat.
  - Linear regression of these prosodic scores to predict human scores.
  - The correlation bet. machine and human is not high.

(b) a Japanese novice learner

… wonder – $\boxed{\text{ful weath}}$ – er …

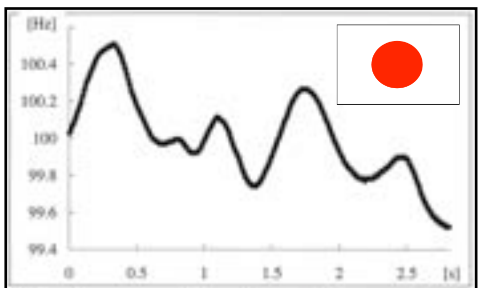
(a) CU based on English syllables

… wonderfu – $\boxed{\text{l wea}}$ – ther …

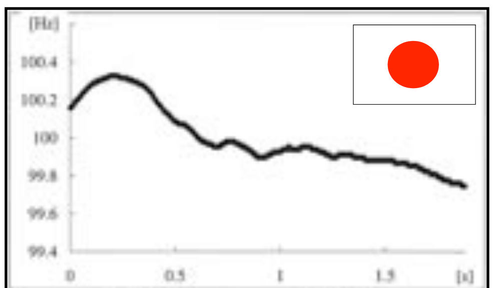
(*waNdafu ru we za-*)

(b) CU based on morae of Japanese-like pronunciation

Fig. 2 Definition of the comparison unit.



b) An F_0 pattern with too large F_0 change of words



(a) An F_0 pattern with flat F_0

Table 4 The correlation between the teachers’ score and automatic scoring.

measure set	closed	open
set-0 (baseline)	0.40	0.41
set-I (proposed)	0.69	0.51

Prosodic comparison with DTW

- Prosodic assessment with word importance factors [Suzuki+'08]
  - Word segmentation is done by forced alignment using an ASR engine.
  - Word-based prosodic comparison between a student and a teacher
    - Ratio of word-based durations, DTW of stress patterns (log energy contour) and DTW of intonation patterns (F0 + log energy contour)
  - Word class importance factor is introduced to improve the performance.
    - A sentence score is obtained as linear combination of the word-based scores.
    - Different words should have different contributions to the final prosodic assessment.
    - DTs are trained so that linear regression errors should be minimized.
      - Leaf-node-dependent linear regressions are used.

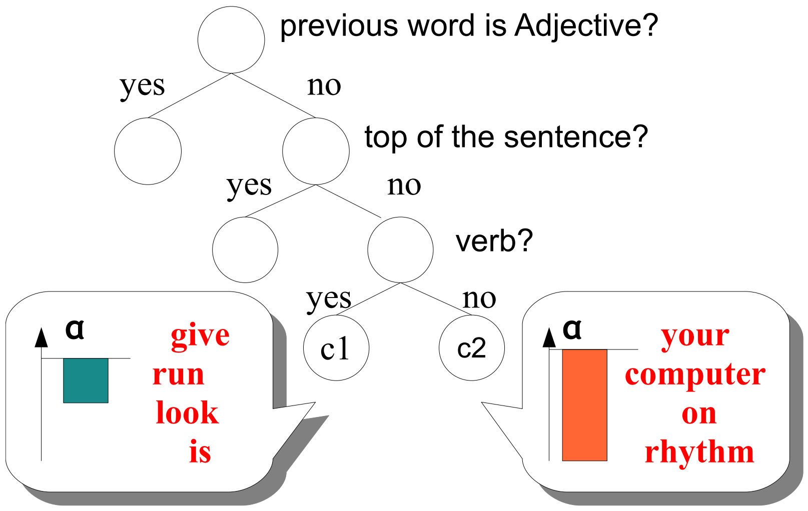
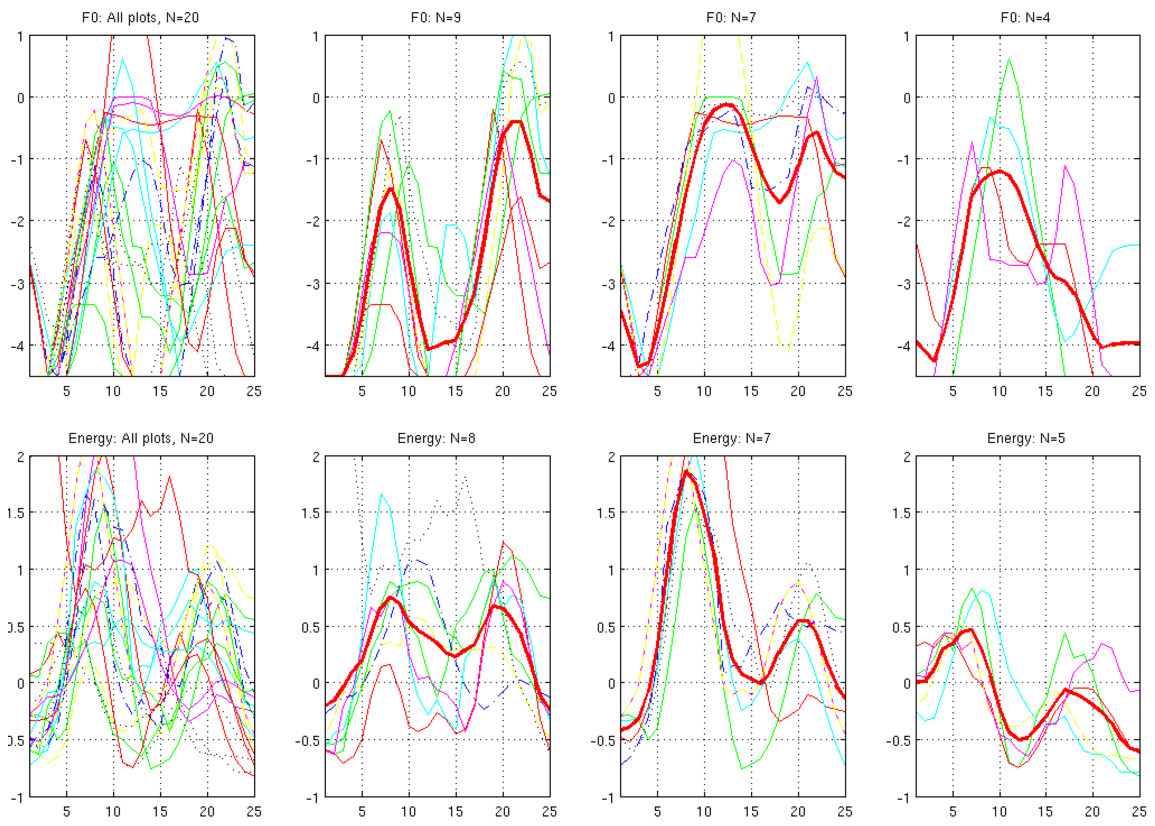


Table 8: Results of intonation evaluation using integration of both scores

	Intonation only	Both scores
Closed	0.59	0.64
Open	0.45	0.48

Prosodic comparison w/o DTW

- Word-based modeling of F0 and energy contours [Cheng'11]
  - 25-point resampling of prosodic patterns for each word
  - Each word has three templates for each of F0 and energy contours.
  - Euclidean distance is used to quantify a difference between a student pattern and a model pattern.



Clustering results of F0 contours and energy contours of word "strategy" using 20 utterances

Prosodic comparison w/o DTW

- Phoneme duration and inter-word silence duration [Cheng'11]
  - Phoneme duration likelihood, similar to [Franco+'00]
    - $\text{log\_seg\_prob} = \frac{1}{N-2} \sum_{i=2}^{N-1} \log(Pr(D_i^{seg}))$
  - Inter-word silence duration likelihood
    - $\text{log\_sil\_prob} = \frac{1}{M} \sum_{i=1}^{M} \log(Pr(D_i^{sil}))$
  - Linear regression of F0, energy, and duration scores to predict human scores

Features	Correlation
F0	0.67
Energy	0.67
F0 + Energy	0.73
$iw\_log\_seg\_prob$	0.54
log_seg_prob	0.76
Linear regression	0.80

Table 2: Correlations using different features.

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL System (TK)
- Database for CALL (NM)

Text-to-speech technology

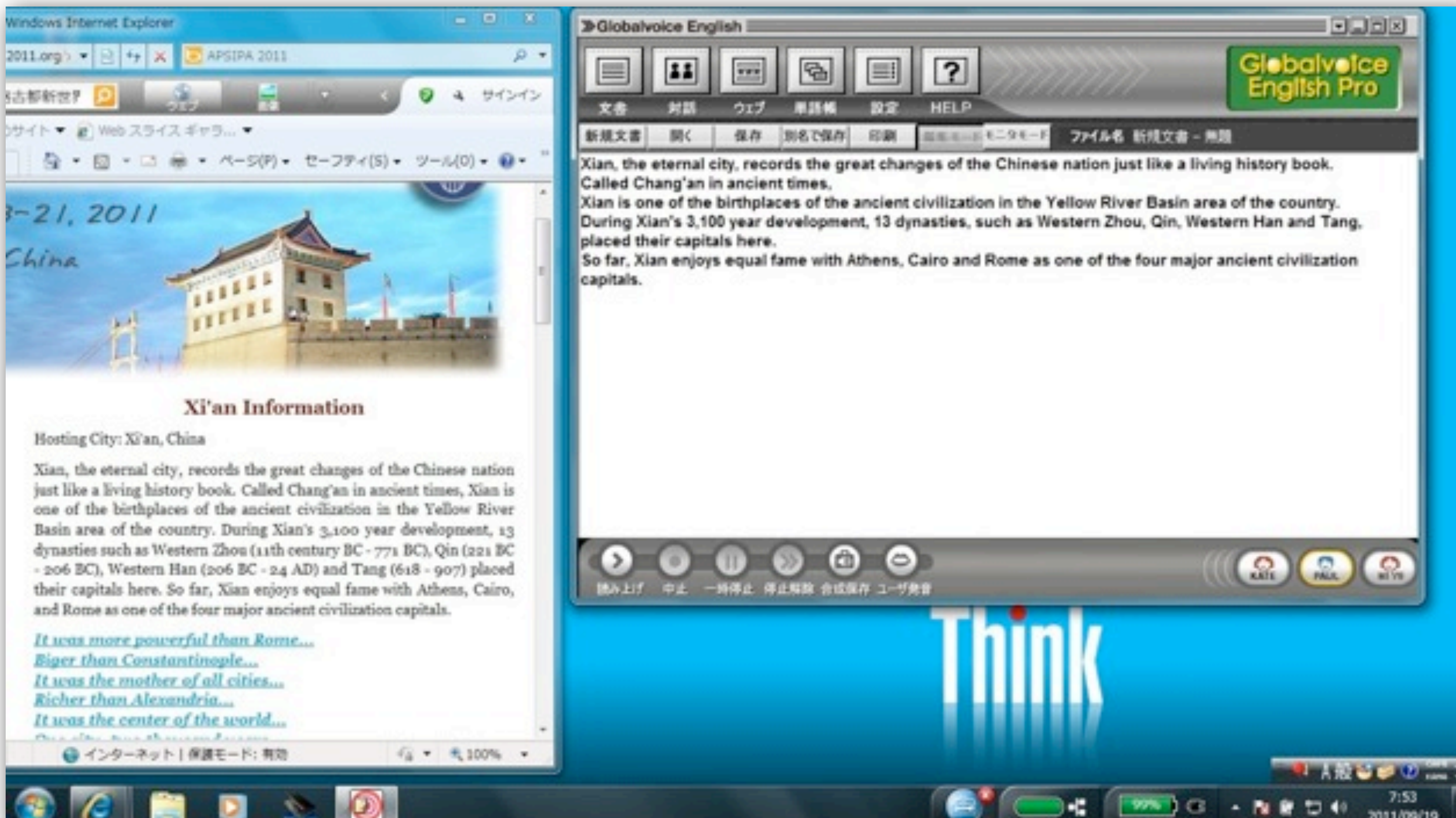
- Two main streams of TTS technology
  - Unit-selection-based generation of waveforms
    - Selection and concatenation of waveform templates
  - HMM-based generation of waveforms
    - Cepstrum-vocoder based generation
  - Comparison of the two frameworks
    - The former tends to be higher in naturalness.
    - The latter is higher in flexible control.

Oxford-Hachette French Dictionary

- Use of TTS technology for CALL [Handley+'05][Black'07]
  - As model pronunciation
    - Use of TTS in pronunciation training
    - Required naturalness is extremely high.
  - As reading machine
    - Use of TTS in dictation practice, shadowing practice, etc
    - Required naturalness is high.
  - As dialogue partner in a dialogue-based CALL system
    - Required naturalness is not so high.

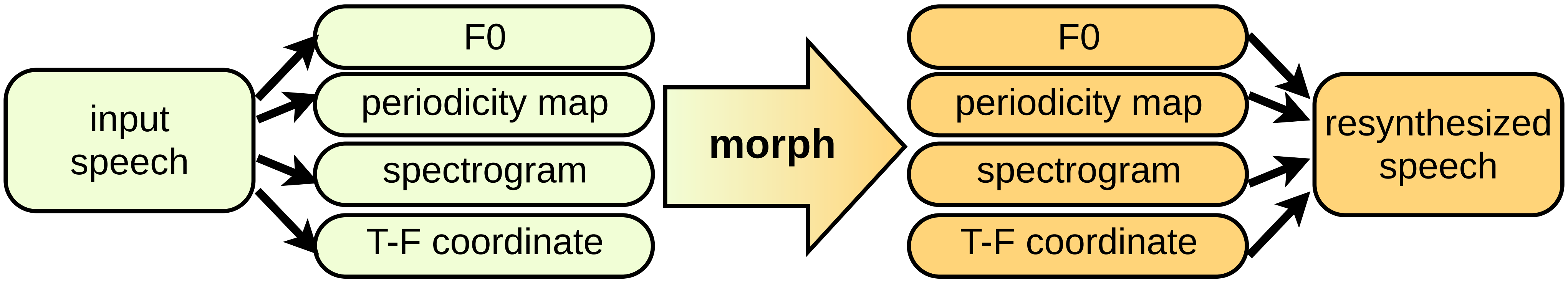
Some demos of high-quality TTS

- "Globalvoice English" produced by HOYA service corp., Japan
  - http://voicetext.jp
  - Used in dictation practice and shadowing practice in college English classes



Use of re-synthesis technology

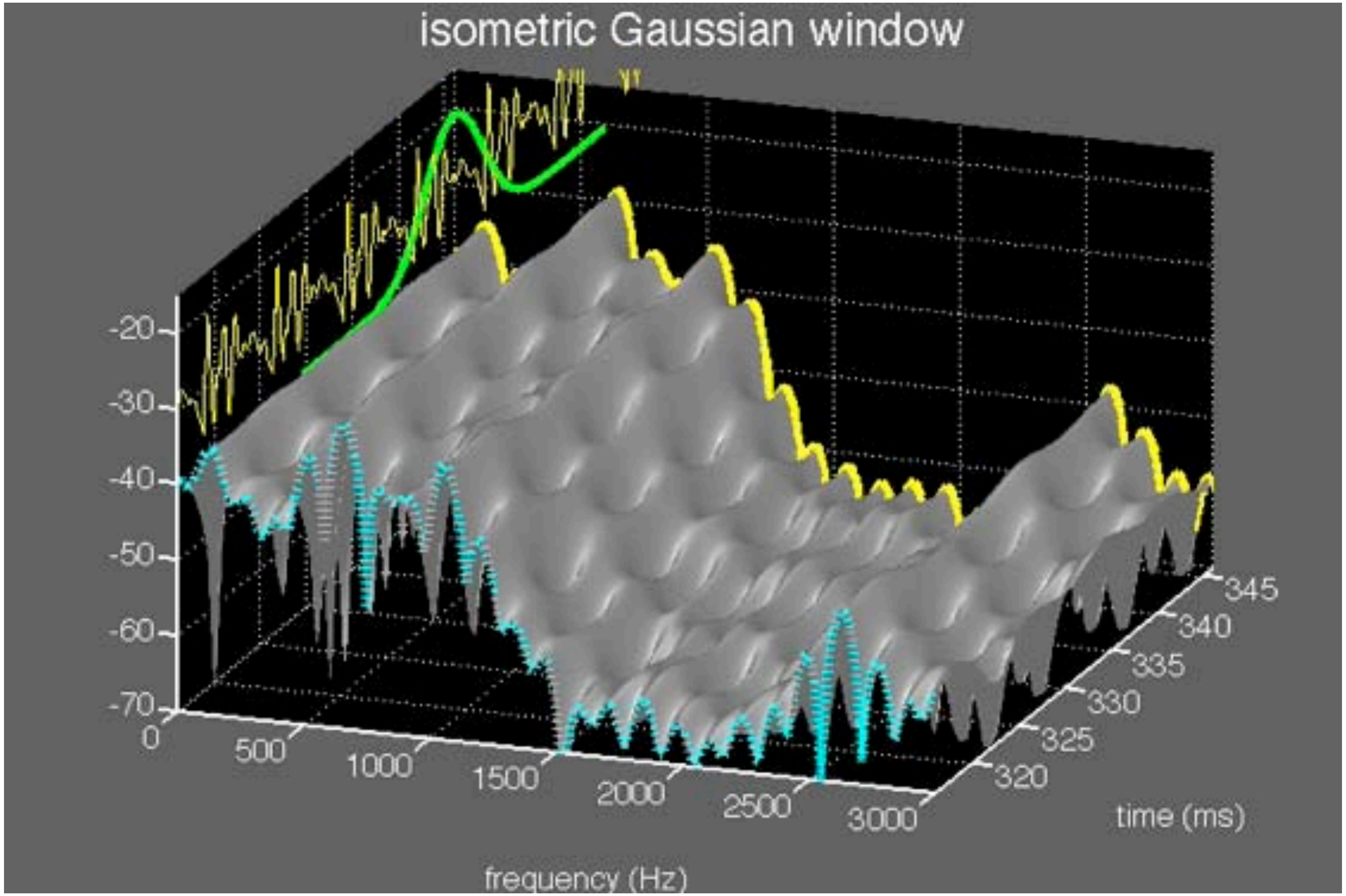
- STRAIGHT [Kawahara'06]
  - High-quality analysis-resynthesis tool
    - Decomposition of speech into
      - Fundamental frequency, spectrographic representations of power, and that of periodicity
    - High-quality speech morphing tool



  - Spectrographic representation of power
    - F0 adaptive complementary set of windows and spline based optimal smoothing
  - Instantaneous frequency based F0 extraction
    - With correlation-based F0 extraction integrated
  - Spectrographic representation of periodicity
    - Harmonic analysis based method

Representation based on SFT

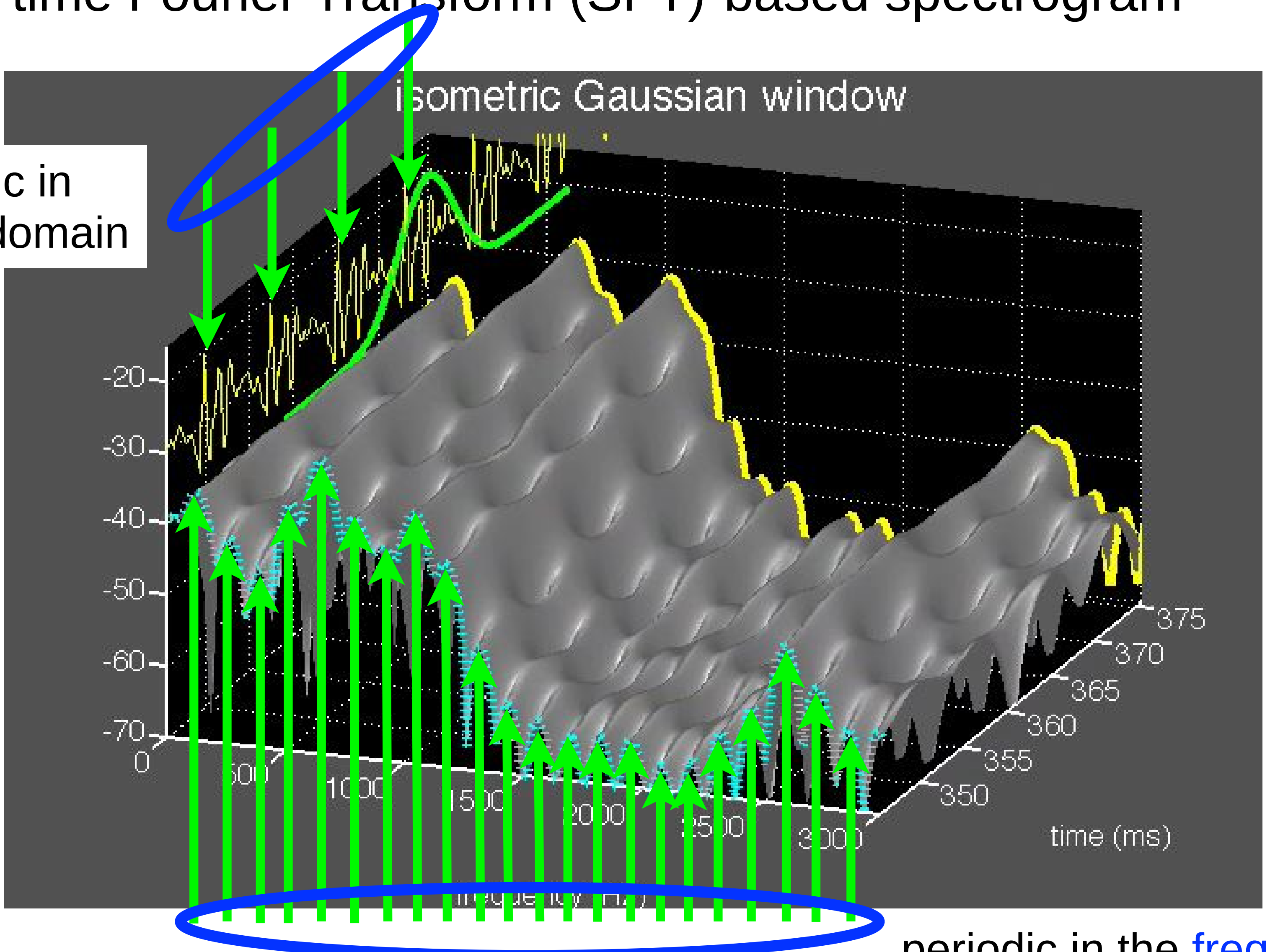
- Short-time Fourier Transform (SFT)-based spectrogram



Representation based on SFT

- Short-time Fourier Transform (SFT)-based spectrogram

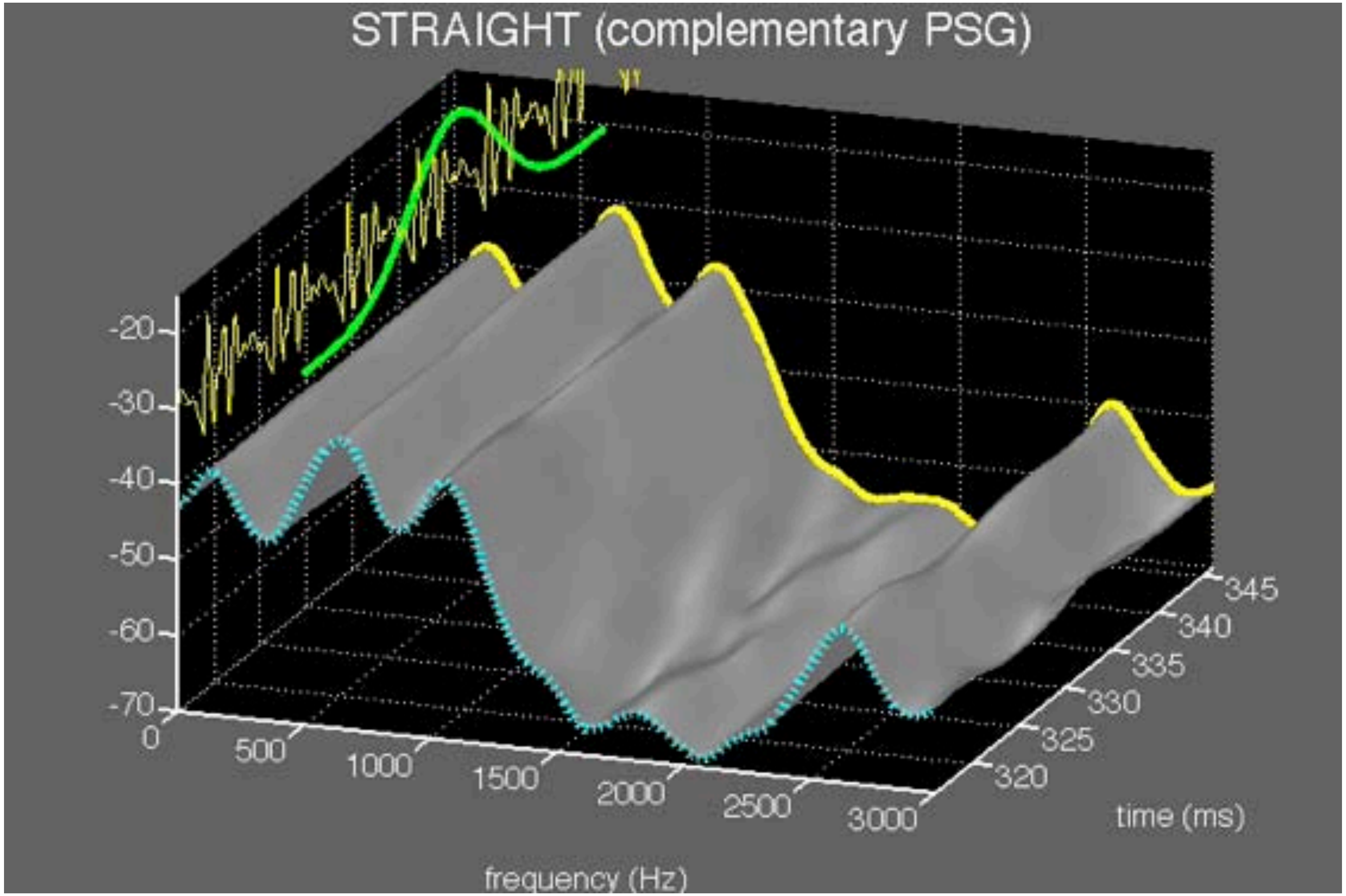
periodic in
the time domain



periodic in the freq. domain

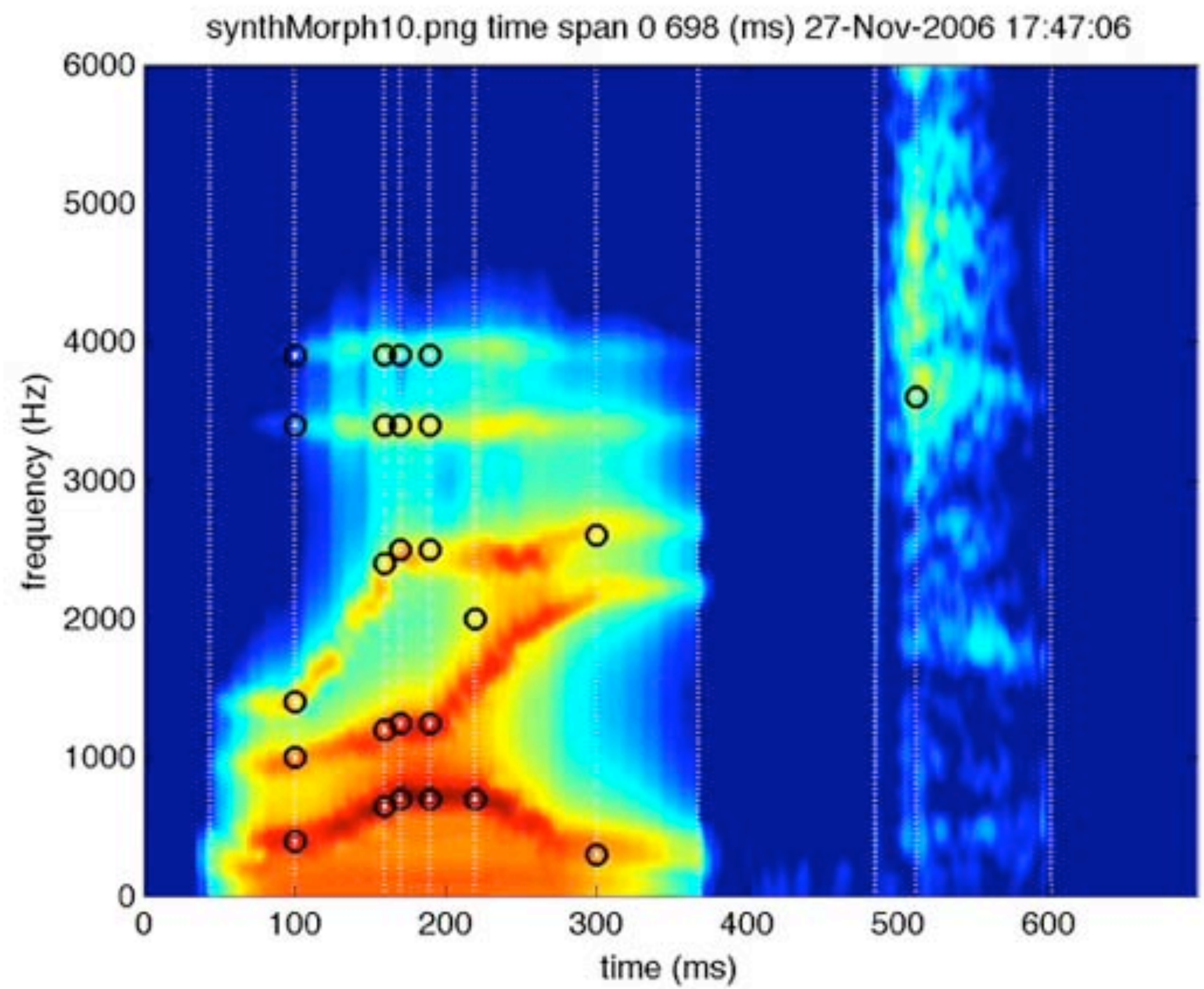
Representation based on STRAIGHT

- Spline-based optimum smoothing reconstructs the underlying smooth time-frequency representation.



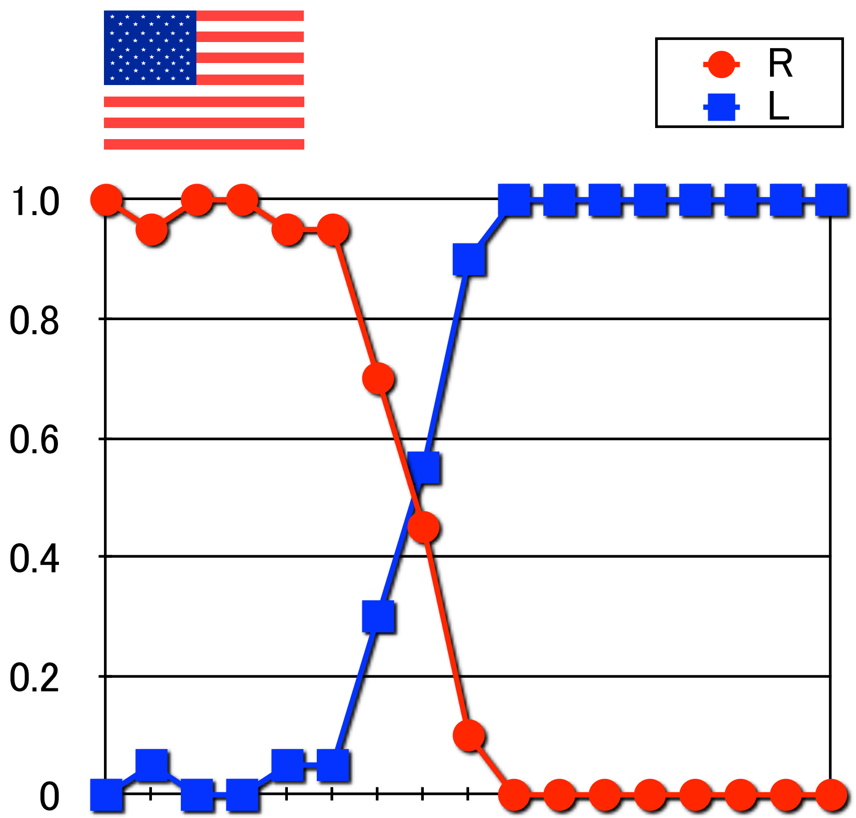
Use of morphed utterances

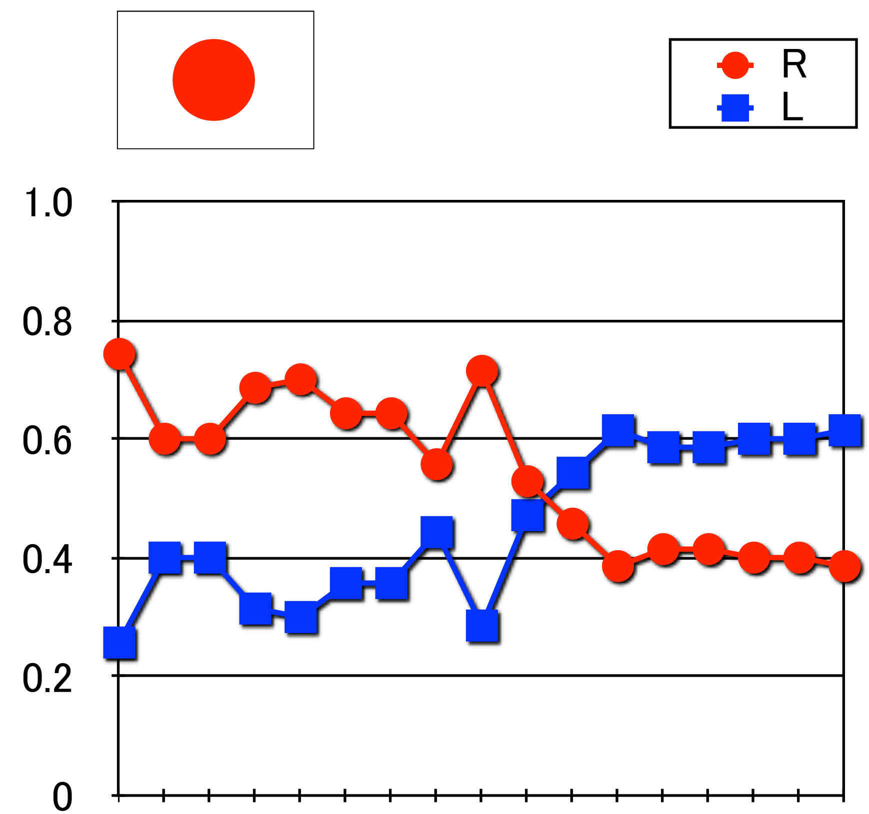
- R to L morphing bet. r/l-ight generated by Klatt synthesizer [Kubo+'98]



Use of morphed utterances

- Results of categorical listening tests [Kubo+'98]
  - 1 American listener
  - 7 Japanese listeners
  - Probability of perceiving R or L in the presented sounds

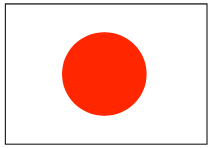




Use of morphed utterances

- Morphing of a native utterance and its accented version [Kato+'11]
  - Use of a pair of word utterances spoken by a bilingual speaker
    - Normal Tokyo Japanese
    - Heavily American accented Japanese

igaku (medical science)



1.

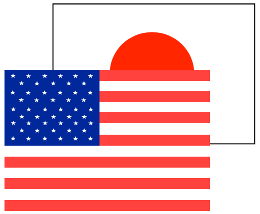
fundamental frequency (F0) → 3.

phonetic duration (dur) → 4.

spectral envelope & aperiodicity (sp_ap) → 5.

F0 & dur (F0_dur) → 6.

all the parameters (all) → 7. = 2.

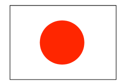
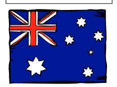


2.

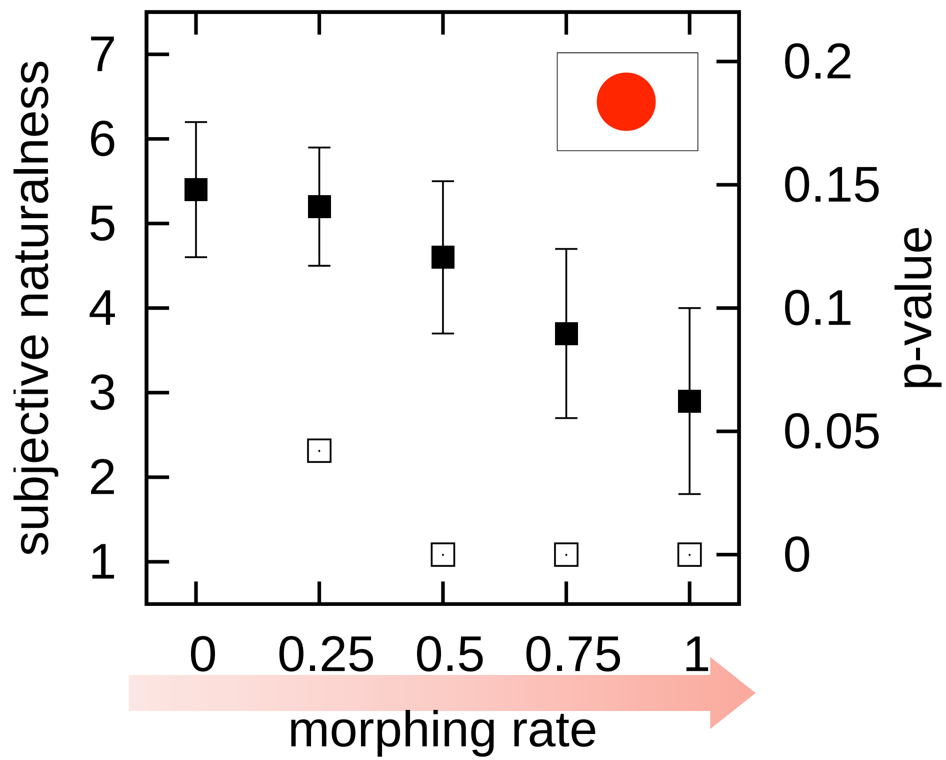
0 0.25 0.5 0.75 1

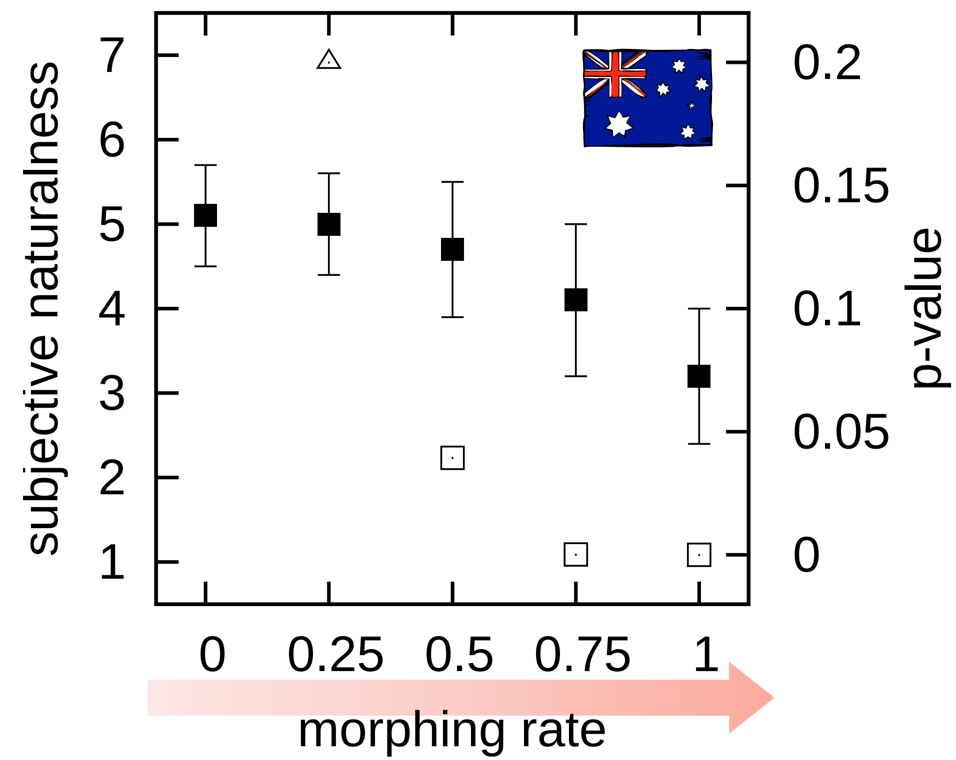
morphing rate

Use of morphed utterances

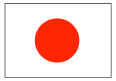
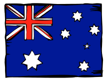
- Prosodic insensitivity of foreign listeners [Kato+'11]
  - 42 Japanese listeners 
  - 15 Australian listeners 
  - Judgement of naturalness as Tokyo Japanese

Morphing only in terms of duration

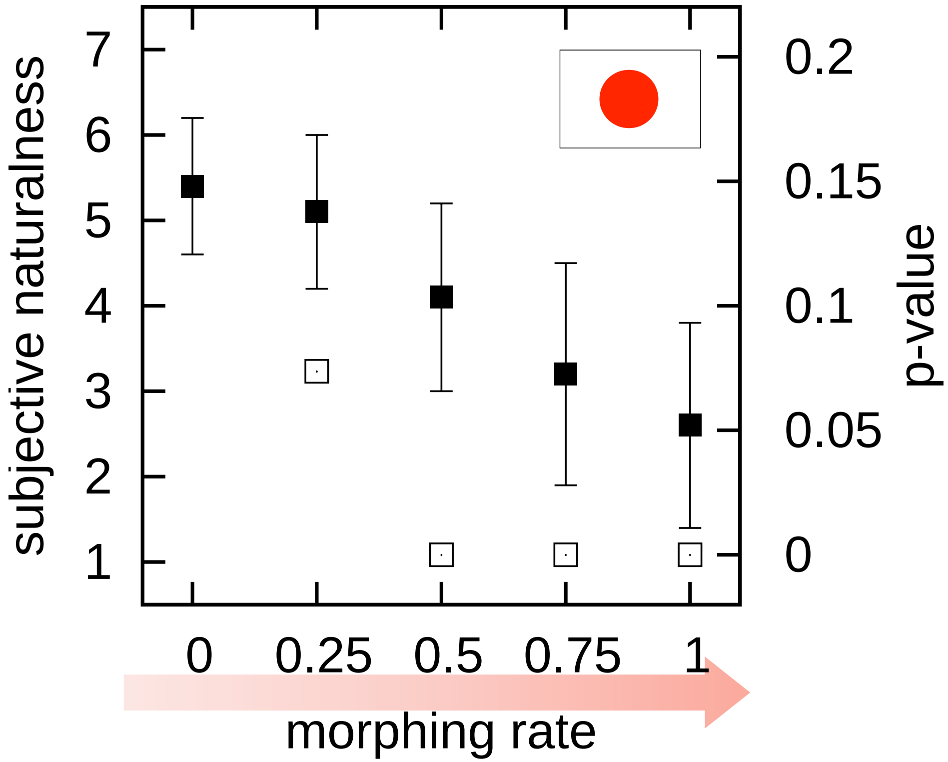


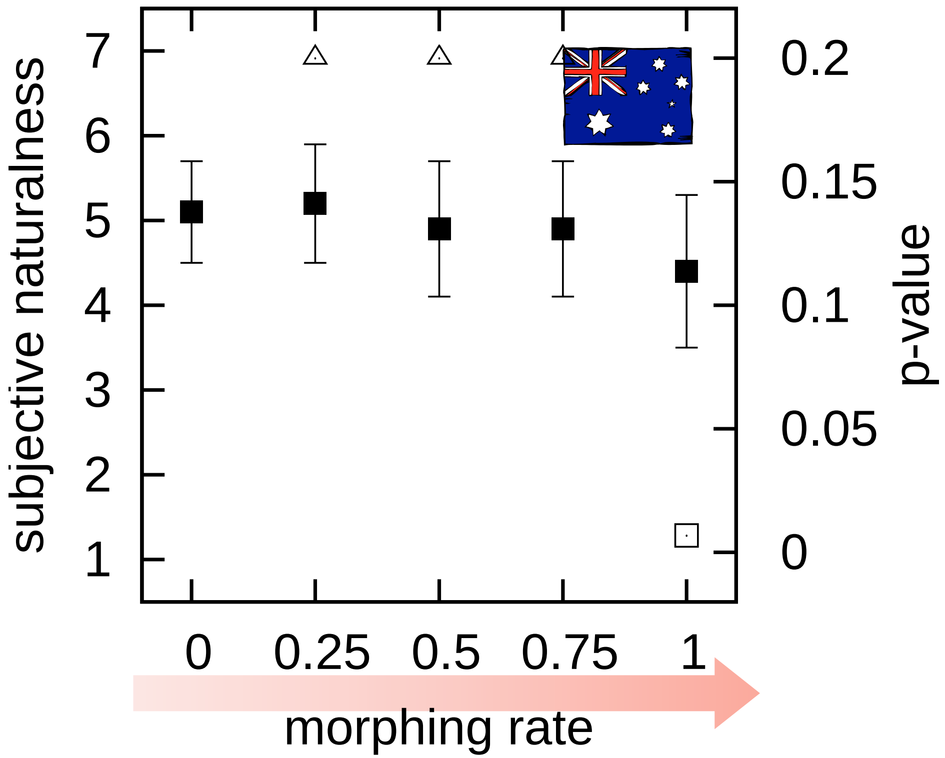


Use of morphed utterances

- Prosodic insensitivity of foreign listeners [Kato+'11]
  - 42 Japanese listeners 
  - 15 Australian listeners 
  - Judgement of naturalness as Tokyo Japanese

Morphing only in terms of F0





Feedback in a learner's own voice

- Prosodic correction of a learner's utterance [Hirose+'03]
  - The corrected version is given to a learner of Japanese as feedback
  - The feedback is generated in his/her own voice.
    - PSOLA (Pitch Synchronous OverLap Add)-based implementation
  - Easy comparison between a bad example and a good one.

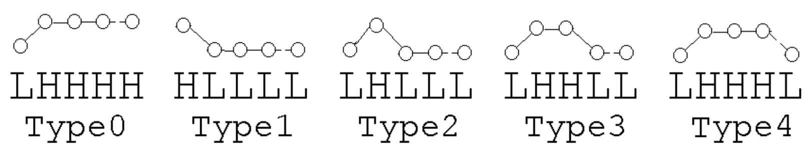


Figure 1: *Binary description of 4-mora Japanese pitch accent patterns. The fifth circle point in each pattern represents pitch level of the attached particle. Type 0 can be distinguished from type 4 by the particle's pitch level.*

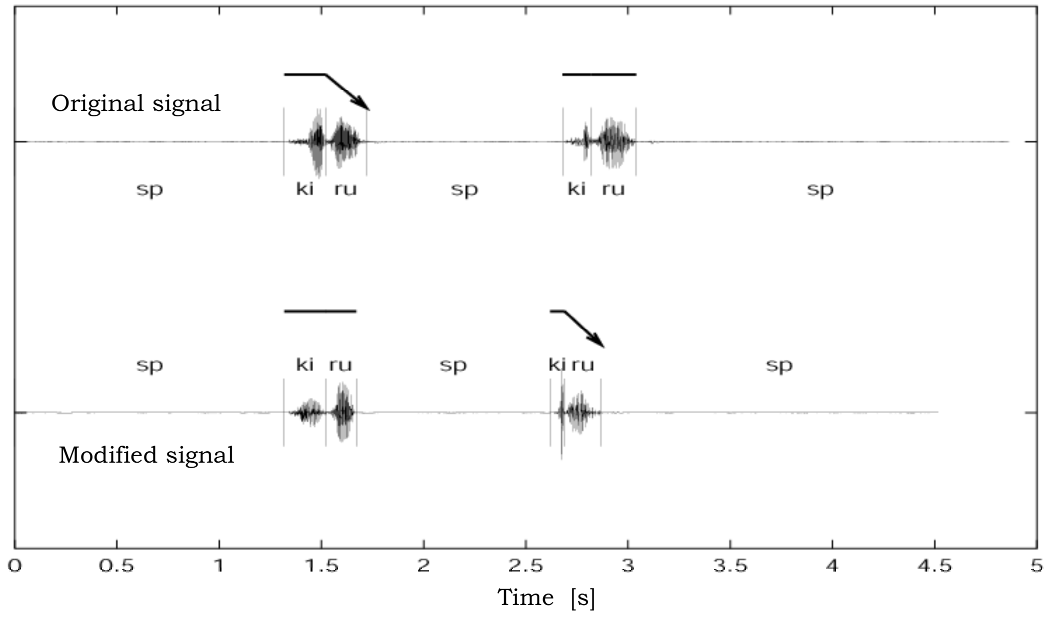


Figure 3: *An example of visual feedback for the couple of homonyms "kiru (to wear)" and "kiru (to cut)."*

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL Systems (TK)
- Database for CALL (NM)

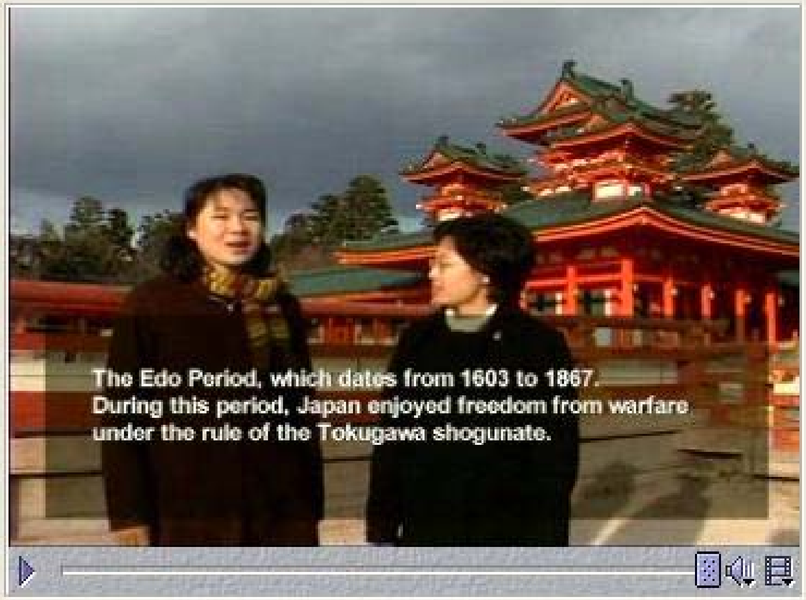
English CALL System: HUGO @Kyoto Univ. [Tsubota, Imoto, Raux 2002]

- For Japanese college students, so that they can introduce Japanese cultures
- Dedicated acoustic model & error prediction scheme for Japanese students
- Deployed and used in classrooms

English CALL System: HUGO @Kyoto Univ.

- Goal: Pinpointing the pronunciation errors which degrade intelligibility and providing effective feedback
- Practice consists of two phases
  1. Dialogue-based skit (for natural conversation)
  2. Training on specific errors detected in the first phase (using a phrase or a word)
- Pronunciation error detection
  - Segmental pronunciation ← hand-crafted phonological rules
  - Accent (Primary & Secondary Stress) ← multiple prosodic features

HUGO Pronunciation Tutor (tsubota) - The Edo Period



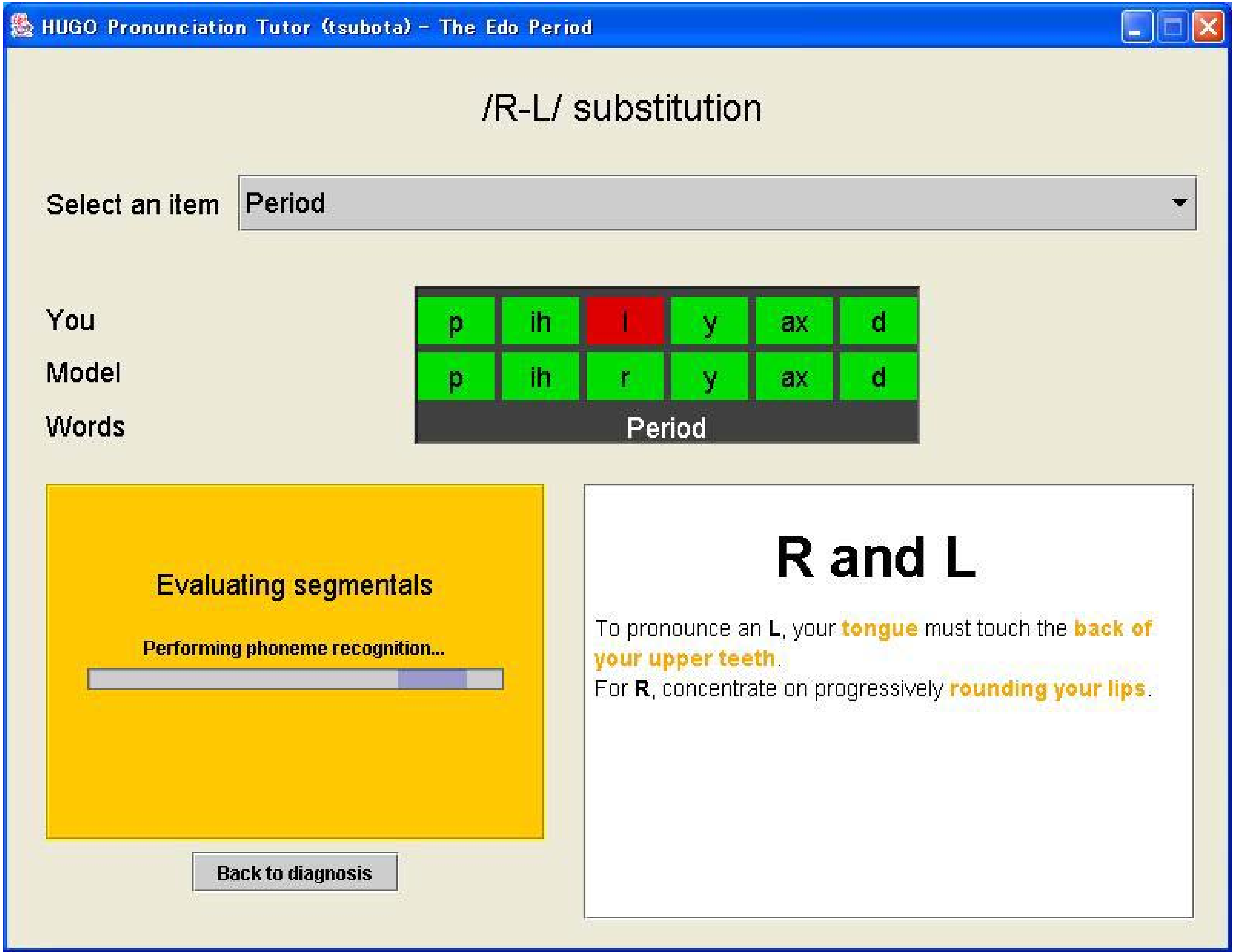
Record

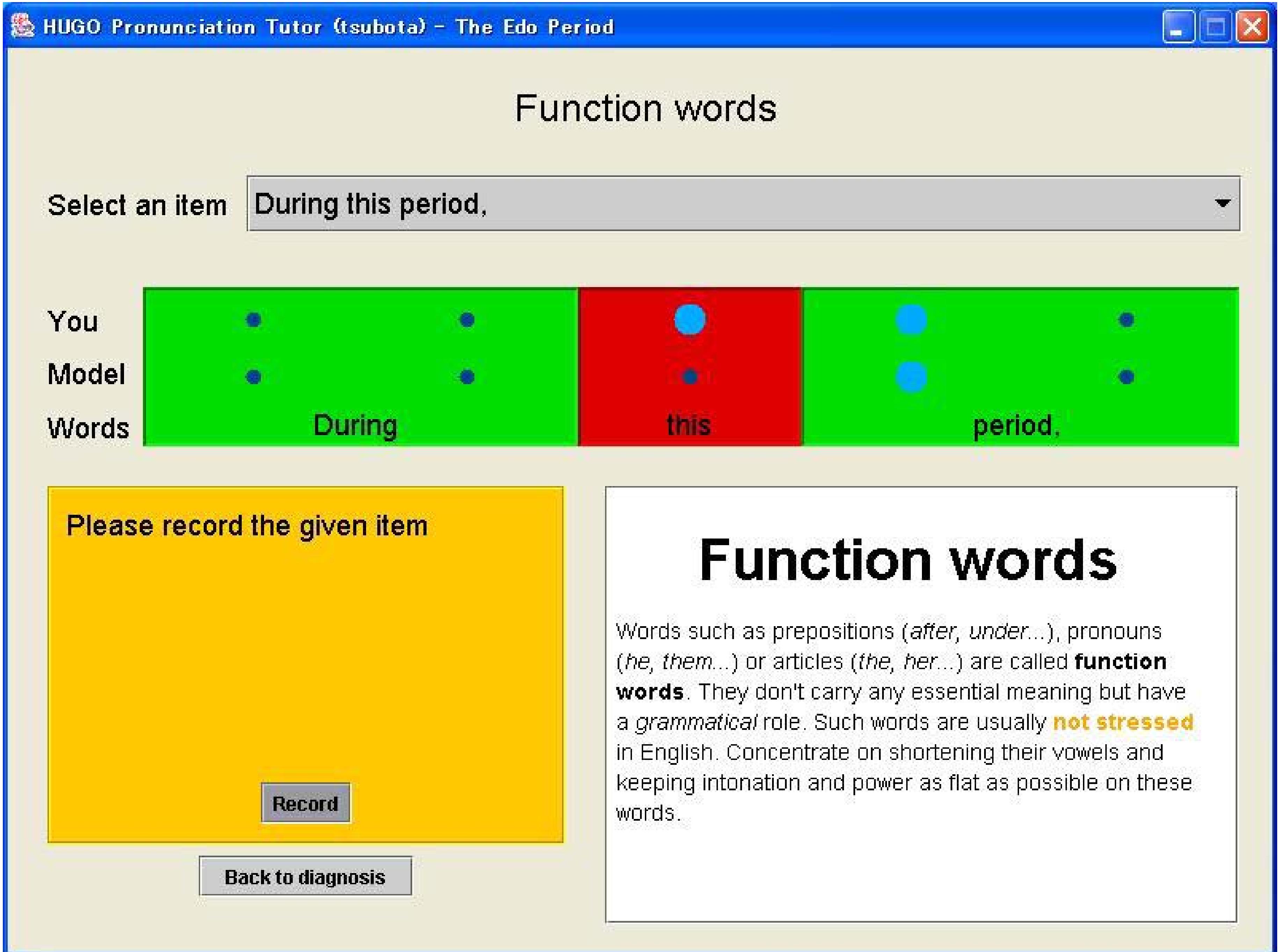
A

What **period** will the next part represent?

B

The Edo Period, which dates from 1603 to 1867. During this period, Japan enjoyed freedom from **warfare** under the rule of the Tokugawa shogunate.


HUGO Pronunciation Tutor (tsubota) - The Edo Period
/R-L/ substitution
Select an item
Period
You
Model
Words
p ih l y ax d
p ih r y ax d
Period
Evaluating segmentals
Performing phoneme recognition...
Back to diagnosis
R and L
To pronounce an L, your tongue must touch the back of your upper teeth.
For R, concentrate on progressively rounding your lips.

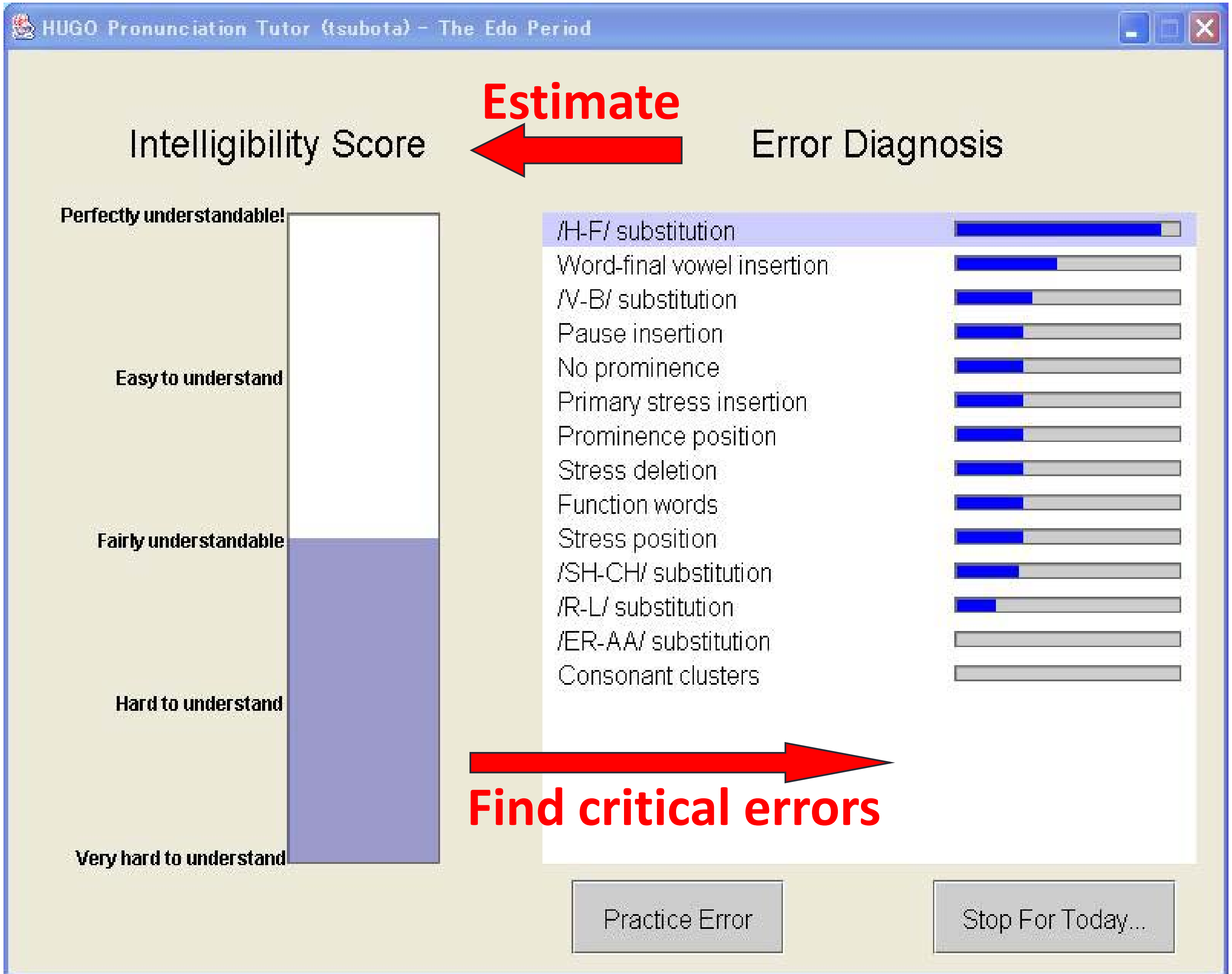

HUGO Pronunciation Tutor (tsubota) - The Edo Period
Function words
Select an item
During this period,
You
Model
Words
During
this
period,
Please record the given item
Record
Back to diagnosis
Function words
Words such as prepositions (*after*, *under*...), pronouns (*he*, *them*...) or articles (*the*, *her*...) are called **function words**. They don't carry any essential meaning but have a *grammatical* role. Such words are usually **not stressed** in English. Concentrate on shortening their vowels and keeping intonation and power as flat as possible on these words.

List of Pronunciation Errors

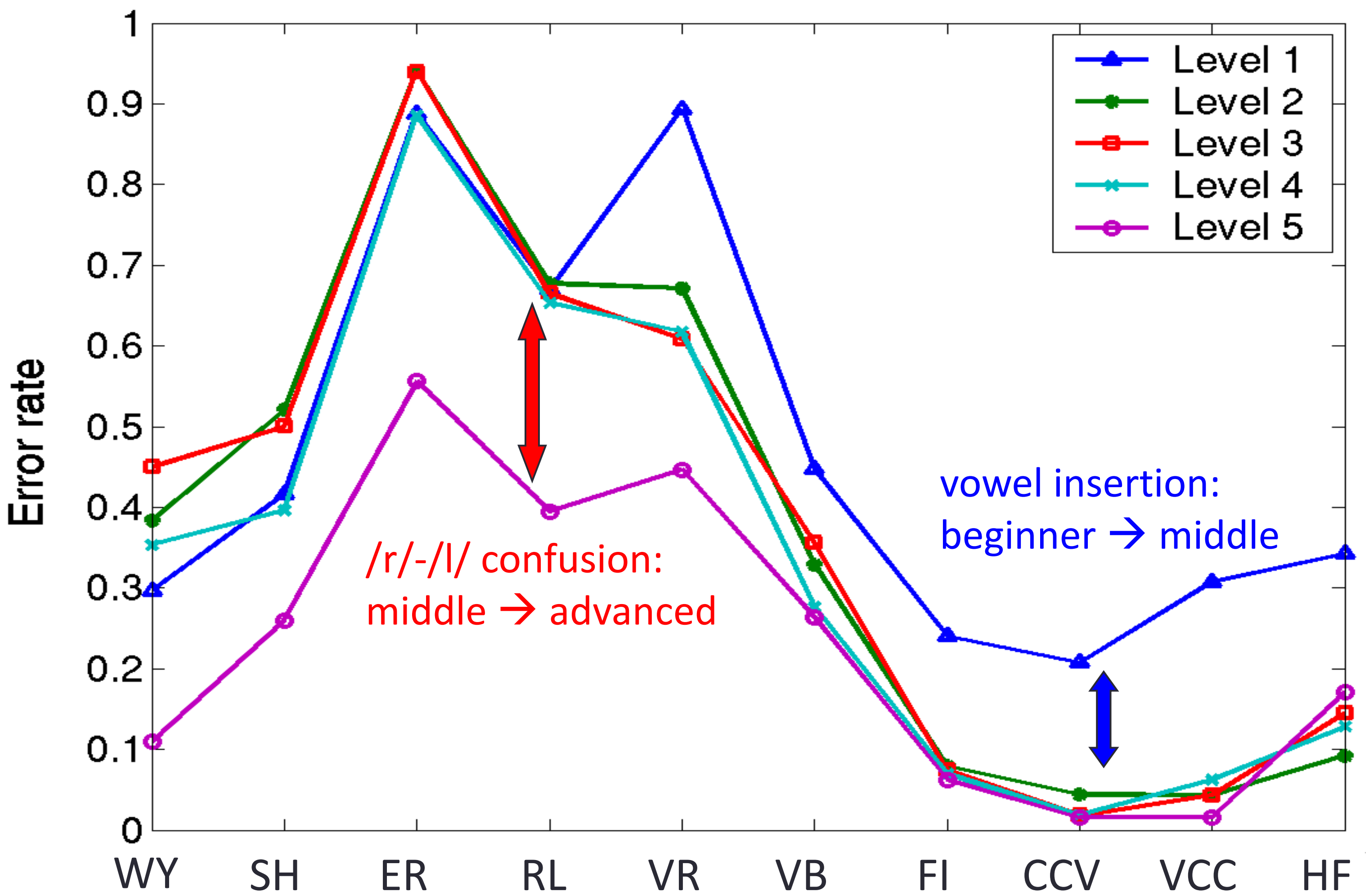
W/Y deletion (would)	V/B substitution (problem)
SH/CH substitution (choose)	Final vowel insertion (let)
R/L substitution (road)	CCV-cluster insertion (active)
ER/A substitution (paper)	VCC-cluster insertion (study)
Non-reduction (student)	H/F substitution (fire)

- Built from literature in ESL
- Remove error patterns with low detection rate

Intelligibility Assessment based on Error Statistics



Priority of Training on Specific Errors according to Intelligibility Level



NativeAccent [Eskenazi 2007]

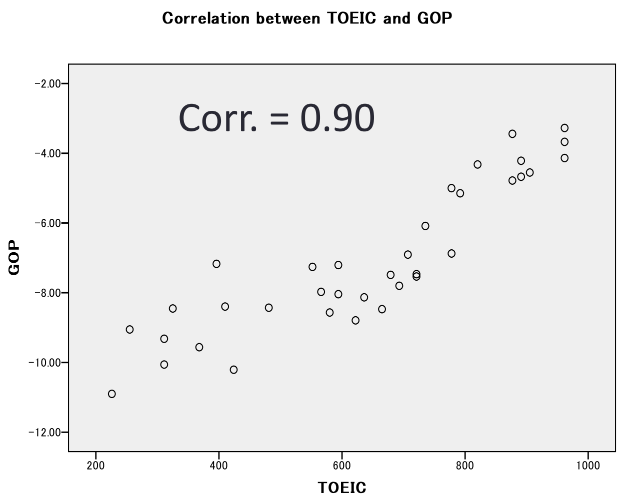
- Product of Fluency Project of CMU
- English learning
  - Error detection and feedback on articulation
  - Up to 28 L1: Japanese, Russian, French…
  - 800 exercises

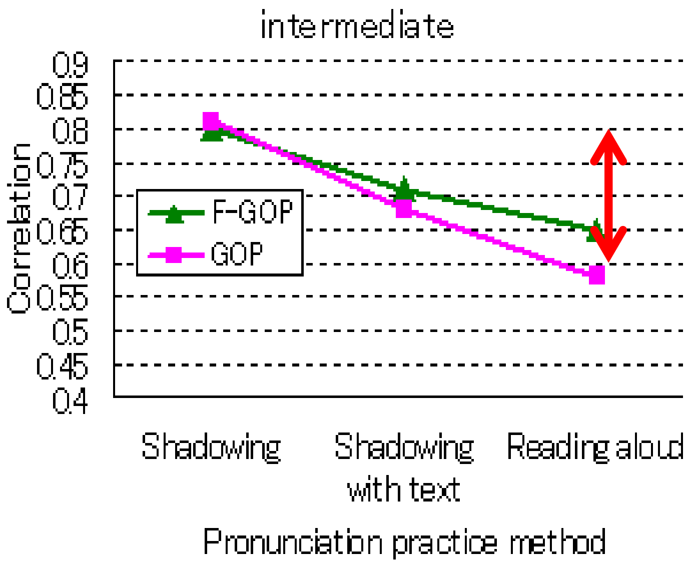
English CALL System @ CUHK [Meng 2010]

- For Chinese learners of English
- Corpus: 100 Cantonese and 111 Mandarin L1
  - Reading a paragraph, words
- Pronunciation error model
  - Hand-crafted phonological rules
  - Data-driven patterns
- GOP score
- Pre-filtering based on duration models
- Synthesizing expressive speech to convey emphasis in feedback generation
- Synthesizing visual speech with articulator animation

Shadowing Exercise [Luo 2009]

- listening and repetition of native utterances, online
  - Simultaneous training of listening and speaking skills
- High correlation between GOP and TOEIC scores (= 0.90)
  - Higher than simple reading





ETS SpeechRater for TOEFL [Zechner 2007]

- Assessment of unconstrained English speech
  - TOEFL iBT Practice Online (TPO)
  - iBT Field Study
- Acoustic model: non-native speech (30hours)
- Language model: non-native speech + broadcast news
- Features: ASR results (word ID, confidence), speech rate, pause length… 40 in total
- Scoring: linear regression model
- Correlation with human rater: 0.67
  - Inter-human correlation 0.94

Dialog-Based English CALL @POSTECH [Lee 2010]

- Situated dialog…(ex.) shopping
- ASR+SLU
- Example-Based Dialog Management
  - very limited domain
- Corrective feedback based on example selection
- Field trial on elementary school

ETS SpeechRater for TOEFL [Zechner 2007]

- Assessment of unconstrained English speech
  - TOEFL iBT Practice Online (TPO)
  - iBT Field Study
- Acoustic model: non-native speech (30hours)
- Language model: non-native speech + broadcast news
- Features: ASR results (word ID, confidence), speech rate, pause length… 40 in total
- Scoring: linear regression model
- Correlation with human rater: 0.67
  - Inter-human correlation 0.94

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL System (TK)
- Database for CALL (NM)

Speech database distribution sites

- Useful information source for speech databases
  - Linguistic Data Consortium (LDC, US)
    - http://www.ldc.upenn.edu/
  - European Language Resources Association (ELRA, EU)
    - http://www.elra.info/
  - Speech Resource Consortium (SRC, Japan)
    - http://research.nii.ac.jp/src/, http://research.nii.ac.jp/src/eng/index.html
  - Advanced LAnGuage INformation forum (ALAGIN, Japan)
    - http://www.alagin.jp/, http://www.alagin.jp/index-e.html
  - GSK (Gengo-Shigen-Kyokai = Langauge Resource Association, Japan)
    - http://www.gsk.or.jp/index.html, http://www.gsk.or.jp/index_e.html
  - Chinese Linguistic Data Consortium (C-LDC, China)
    - http://www.chineseldc.org/

  - These sites distribute speech & language databases for general purposes.
  - Only a part of the databases include non-native speech samples.

Non-native speech data collection

- More useful information source for non-native speech data
  - "Non-native speech database" in Wikipedia
    - http://en.wikipedia.org/wiki/Non-native_speech_database
    - Based on [M. Raab+'07]
    - 42+ non-native databases are briefly described.

Corpus	Author	Available at	Language(s)	#Speakers	native Language	#Utt.	Duration	Date	Specials	Reference
AMI		EU	E		Dut and other		100h		meeting recordings	#40
ATR-Gruhn	Gruhn	ATR	E	96	C G F J Ind	15000		2004	proficiency rating	#4
BAS Strange Corpus I+II		ELRA	G	139	50 countries	7500		1998		#5
Berkeley Restaurant		ICSI	E	55	G I H C F S J	2500		1994		#41
Broadcast News		LDC	E					1997		#6
Cambridge-Witt	Witt	U. Cambridge	E	10	J I K S	1200		1999		#7
Cambridge-Ye	Ye	U. Cambridge	E	20	C	1600		2005		#8
Children News	Tomokiyo	CMU	E	62	J C	7500		2000	partly spontaneous	#6
ERJ	Minematsu	U. Tokyo	E	200	J	68000		2002	proficiency rating	#13

Development of ERJ database

- ERJ = English Read by Japanese [Minematsu+'04]
  - Development of a database containing many pronunciation errors that are observed commonly in the English spoken by Japanese
  - A main focus is put on the errors that are made rather unconsciously.
  - Spontaneous speech is technically challenging. So read speech is focused on.
  - Target language = General American English (GAE)
- Selection of reading material
  - Word and sentence sets considering the segmental aspects of GAE
  - Word and sentence sets considering the prosodic aspects of GAE
  - In total, 807 sentences and 1009 words are prepared.

Table 1: Word and sentence sets for the segmental aspect

set	size
Phonemically-balanced words	300
Minimal pair words	600
TIMIT-based phonemically-balanced sentences	460
Sentences including phoneme sequences difficult for Japanese to pronounce correctly	32
Sentences designed for test set	100

Table 2: Word and sentence sets for the prosodic aspect

set	size
Words with various lexical accent patters	109
Sentences with various intonation patterns	94
Sentences with various rhythm patterns	121

Development of ERJ database

- Preparation of reading sheets
  - Many pronunciation guides are on the sheets
    - Phonemic symbols
    - Stress marks
    - Intonation curves
    - etc.

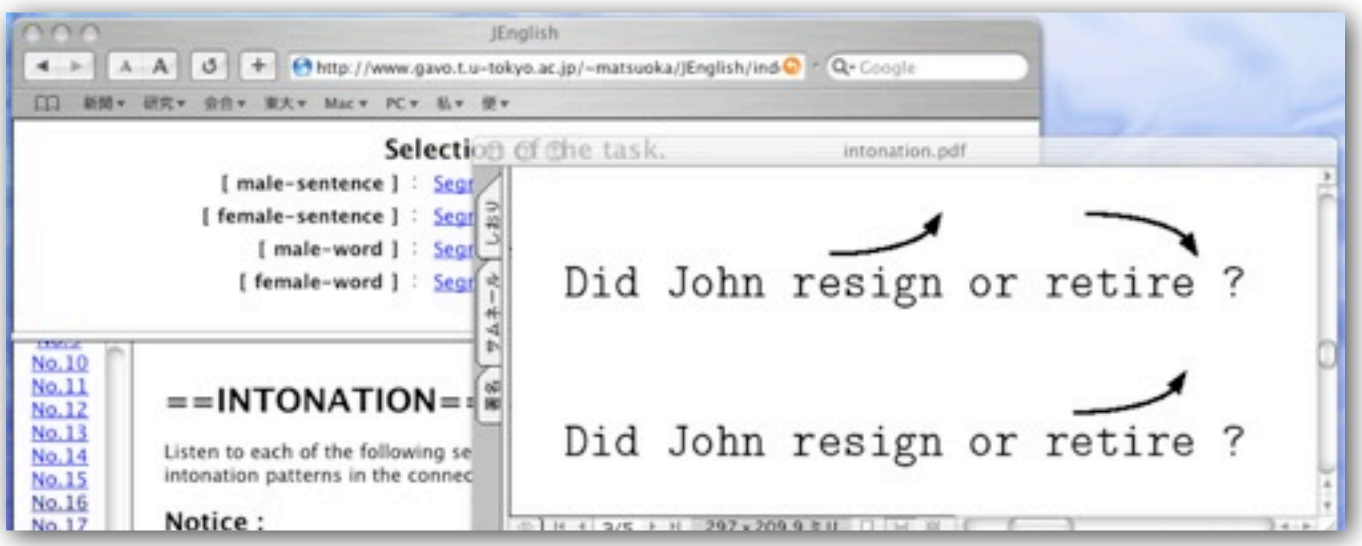
```
S1_0097
          She knows you, doesn't she ?
          [SH IY1] [N OW1 Z] [Y UW1] [D AH1 Z AX0 N T] [SH IY1]

S1_0105   Come to tea.
         / +    -  @ /
          [K AH1 M] [T UW1] [T IY1]
S1_0106   Come to tea with John.
         / +    -  +   -    @  /
          [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N]
S1_0107   Come to tea with John and Mary.
         / +    -  @ / -    +   -    @ -/
          [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N] [AE1 N D] [M EH1 R IY0]
```

Development of ERJ database

- Selection of speakers
  - Quasi-random selection of university/college students of Japanese
  - 100 male and 100 female Japanese
  - 20 General American English (GAE) speakers
- Recording protocol
  - About 120 sentences and 220 words are assigned to each student.
    - About 400 sentences are assigned to each of the 20 American speakers.
  - Pronunciation guides are shown in the reading sheet.
  - The speakers read the material repeatedly until they “thought” that they read the material correctly.
    - *Error-free* utterances judged by the speakers themselves.
    - Still, many errors can be detected by teachers.

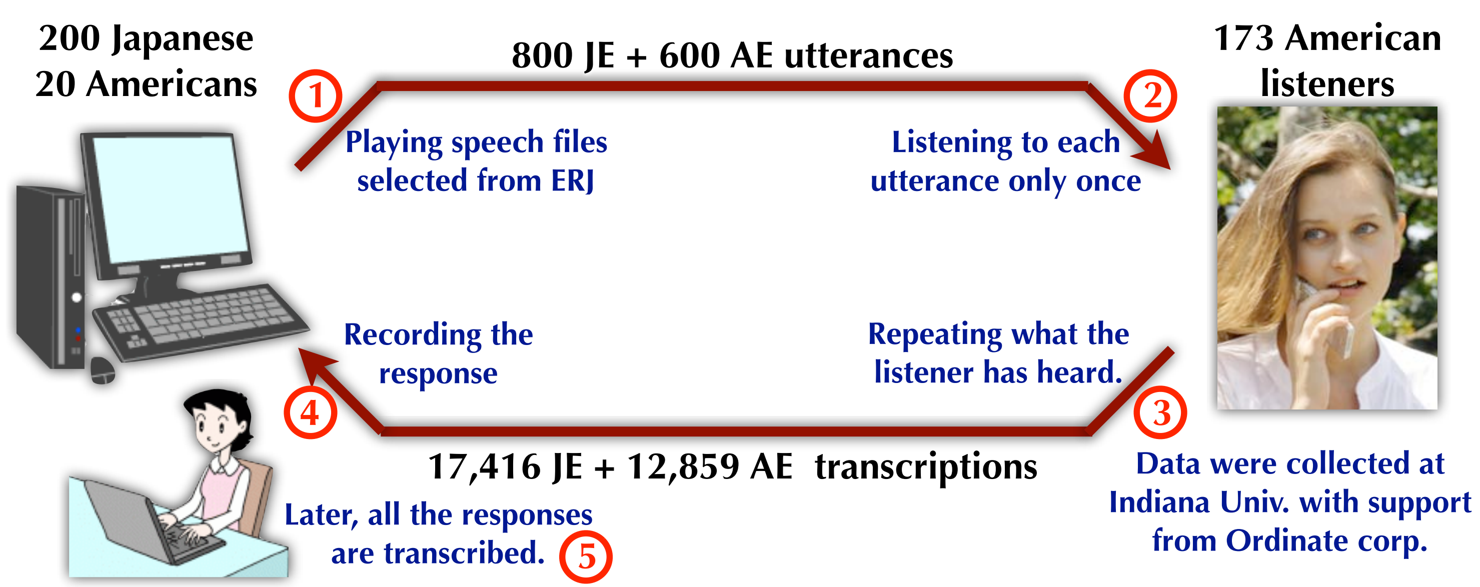
Development of ERJ database

- Rating protocol
  - Five American teachers of English are asked to rate some utterances of the individual students w.r.t the three aspects of pronunciation.
    - Phonemic aspect / intonational aspect / rhythmic aspect
    - As for prosodic rating, model utterances were presented to the teachers because they claimed that the task was difficult without prosodically *perfect* utterances.
- Use of the database
  - Development of CALL systems and their modules
  - Acoustic analysis of Japanese English



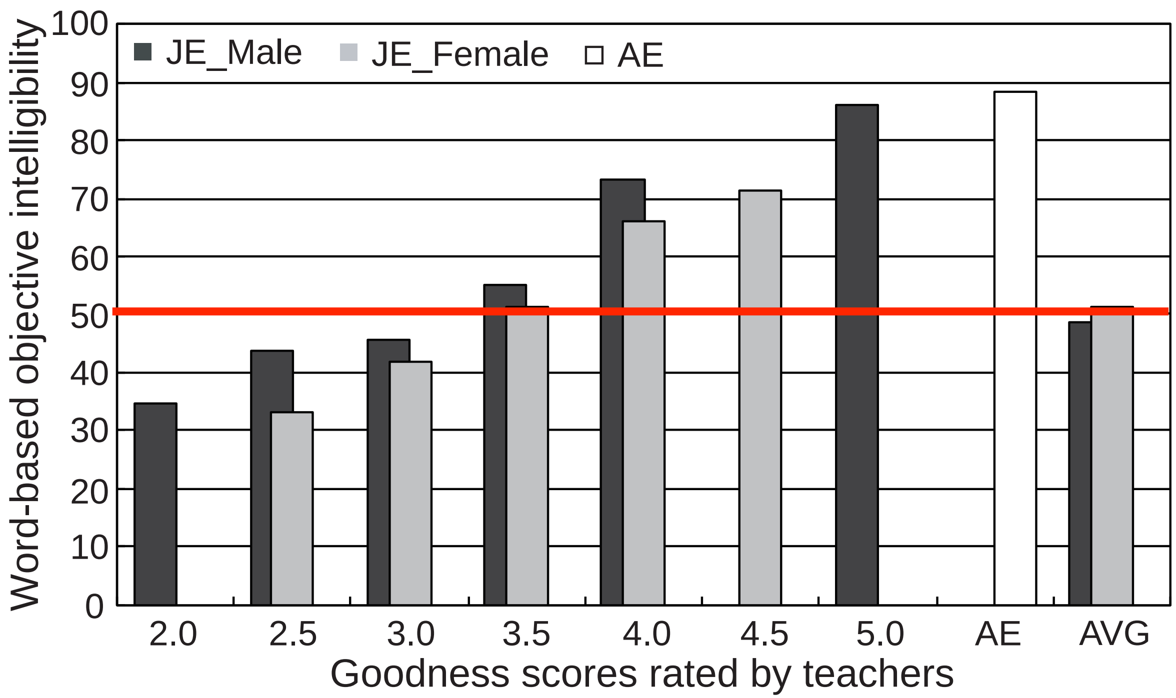
Objective measurement of intelligibility

- How intelligible is JE? [Minematsu+'11]
  - ERJ = many read utterances judged as *error-free* by the students
    - Are these utterances understood correctly by US people?
  - A huge listening test was done using a subset of ERJ database.
    - Listeners : American with little exposure to Japanese English.
      - JE utterances are presented through a telephone line.
    - Task : just repeating what they have heard without trying to guess.
      - Presentation of each utterance was done only once.
      - Repetitive responses were transcribed by expert transcribers.



Objective measurement of intelligibility

- How intelligible is JE? [Minematsu+'11]
  - ERJ = many read utterances judged as *error-free* by the students
    - Are these utterances understood correctly by US people?

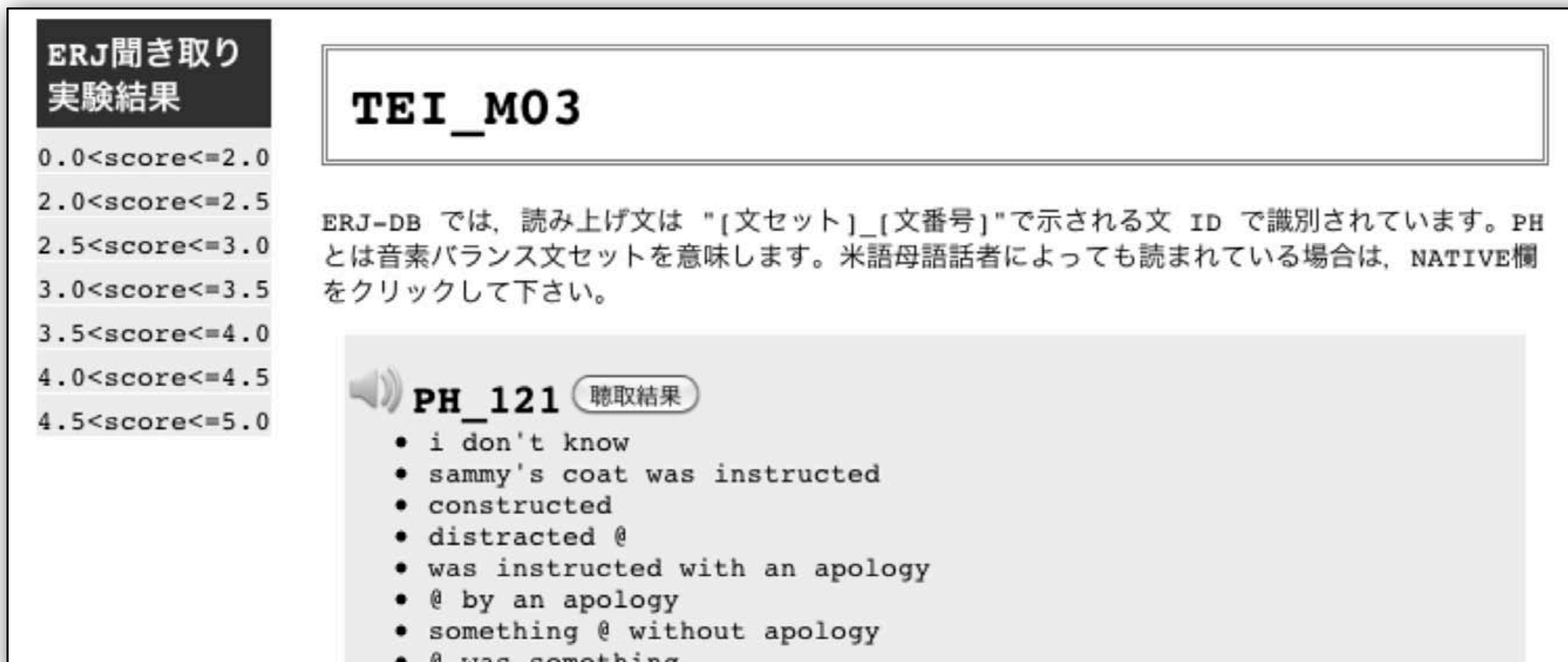


Classification of speakers based on their proficiency scores

score	≤2.0	≤2.5	≤3.0	≤3.5	≤4.0	≤4.5	≤5.0
male	2	27	43	16	5	0	2
female	0	8	36	25	19	7	0

Objective measurement of intelligibility

- Transcription browser [Minematsu+'11]
  - Many *facts* of miscommunication
  - All the utterances used in the large listening test and their transcriptions will be added to the next release of ERJ database.
  - A browsing system for the utterances/transcriptions will be included.
    - #transcription per utterance is 21 on average.



Non-native speech data collection

- More useful information source for non-native speech data
  - Non-native database in Wikipedia
    - http://en.wikipedia.org/wiki/Non-native_speech_database
    - Based on [M. Raab+'07]
    - 42+ non-native databases are briefly described.

Corpus	Author	Available at	Language(s)	#Speakers	native Language	#Utt.	Duration	Date	Specials	Reference
AMI		EU	E		Dut and other		100h		meeting recordings	#40
ATR-Gruhn	Gruhn	ATR	E	96	C G F J Ind	15000		2004	proficiency rating	#4
BAS Strange Corpus I+II		ELRA	G	139	50 countries	7500		1998		#5
Berkeley Restaurant		ICSI	E	55	G I H C F S J	2500		1994		#41
Broadcast News		LDC	E					1997		#6
ERJ	Minematsu	U. Tokyo	E	200	J	68000		2002	proficiency rating	#13

- Data collection is a tough work.
  - Resource sharing is very important.

OUTLINE

- Introduction (TK)
- Segmental Aspect & Speech Recognition Tech. (TK)
  - Pronunciation Structure Model (NM)
- Prosodic Aspect (NM)
- Speech Synthesis Tech. for CALL (NM)
- CALL System (TK)
- Database for CALL (NM)

References

- M. Eskenazi, “An overview of spoken language technology for education,” Speech Communication, 51, 832-844, 2009
- Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., Barker, P., “Applications of automatic speech recognition to speech and language development in young children,” Proc. ICSLP, 1996
- Bernstein, J. Intelligibility and simulated deaf-like segmental and timing errors. Proc. ICASSP, 244-247, 1977
- L. Neumeyer, H. Franco, V. Digalakis and M. Weintraub, “Automatic Scoring of Pronunciation Quality,” Speech Communication, 30, 83-93, 2000
- S. M. Witt and S. J. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,” Speech Communication, 30, 95-108, 2000
- Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., “Automatic evaluation and training in English pronunciation,” Proc ICSLP, 1185-1188, 1990
- N. Minematsu, "Are learners myna birds to the averated distributions of native speakers? -- a note of warning from a serious speech engineer --," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), 2007
- Y. Qiao, N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," IEEE Trans. on Signal Processing, 58, 7, 3884-3890, 2010
- N. Minematsu, "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances," Proc. ICSLP, 1669-1672, 2004
- N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for CALL," Int. Workshop on Spoken Language Technology (SLT), 126-129, 2006
- M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Integration of multilayer regression with structure-based pronunciation assessment," Proc. INTERSPEECH, 586-589, 2010
- M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Pronunciation proficiency estimation based on multilayer regression analysis using speaker-independent structural features," Proc. Int. Workshop on Second Language Studies, 2010
- N. Minematsu, K. Kamata, S. Asakawa, T. Makino, and K. Hirose, "Structural representation of the pronunciation and its use for classifying Japanese learners of English," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), 2007

References

- C. Cucchiarini, H. Strik, L. Boves, "Quantitative assessment of second language leaners' fluency: an automatic approach," Proc. ICSLP, 1998
- C. Cucchiarini, H. Strik. L. Boves, "Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech," J. Acoust. Soc. Am. 111, 6, 2862-2873, 2002
- A. Maier, F. Honig, V. Zeissler, A. Batliner, E. Korner, N. Yamanaka, P. Ackermann, E. Noth, "A language-independent feature set for the automatic evaluation of prosody," Proc. INTERSPEECH, 600-603, 2009
- S. Huang, H. Li, S. Wang, J. Liang. B. Xu, "Automatic reference independent evaluation of prosody quality using multiple knowledge fusions," Proc. INTERSPEECH, 610-613, 2010
- Y. Kim, H. Franco, L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," Proc. EUROSPEECH, 645-648, 1997
- H. Franco, L. Neumeyer, V. Digalakis, O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," Speech Communication, 30, 121-130, 2000
- K. Hirabayashi, S. Nakagawa, "Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques," Proc. INTERSPEECH, 598-561, 2010
- N. Minematsu, N. Ohashi and S. Nakagawa, "Automatic detection of accent in English words spoken by Japanese students," Proc. EUROSPEECH, 701-705, 1997
- N. Minematsu and S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," Proc. Int. Workshop on Integration of Speech Technology into Learning (InSTiLL), 130-137, 2000
- F. Ramus, M. Nespor, J. Mehler, "Correlates of linguistic rhythm in the speech signal," Cognition, 73, 3, 265-292, 1999
- F. Ramus, "Autoamtic correlates of linguistic rhythm: Perspective," Proc. Speech Prosody, 2002
- E. Grabe, B. Post, I. Watson, "The acquisition of rhythmic patterns in English and French," Proc. ICPhS, 1201-1204, 1999
- E. Grabe, E. L. Low, "Durational variability in speech and the rhythm class hypothesis," Laboratory Phonology 7, edited by C. Gussenhoven and N. Warner, Berlin, New York (Mouton de Gruyter), 515-546, 2002
- M. Suzuki, T. Konno, A. Ito, S. Makino, "Automatic evaluation system of English prosody based on word importance factor," J. Systemics, Cybernetics, and Informatics, 6, 4, 83-90, 2008
- J. Cheng, "Automatic assessment of prosody in high-stakes English tests," Proc. INTERSPEECH, 1589-1592, 2011

References

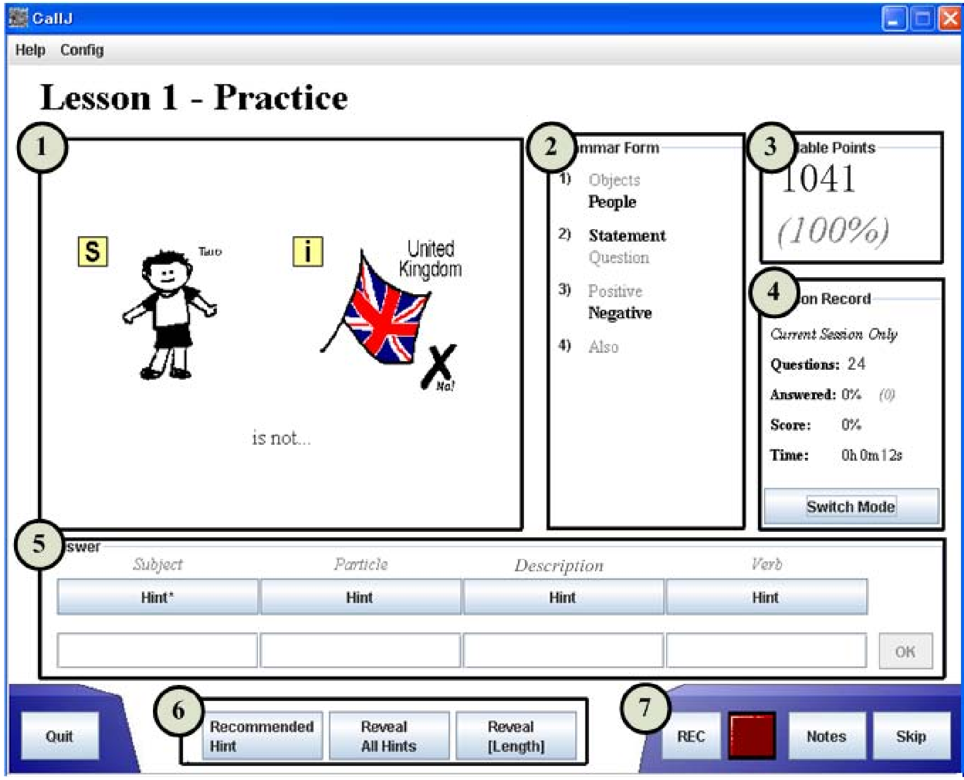
- Y. Yamashita, K. Kato, K. Nozawa, "Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison," IEICE Trans. Inf. & Syst. E88-D, 3, 496-501, 2005
- H-C. Liao, J-C. Chen, S-C. Chang, Y-H. Guan, C-H. Lee, "Decision tree based tone modeling with corrective feedbacks for automatic Mandarin tone assessment," Proc. INTERSPEECH, 602-605, 2010
- K. Hirose, F. Gendrin, N. Minematsu, "A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice," Proc. EUROSPEECH, 3149-3152, 2003
- M. Sugito, "Nihon-jin no Eigo (English pronunciation of Japanese speakers)", Izumishoin, 1998
- K. Shimizu, "Eigo onseigaku (Phonetics of English)", Sokei-shobo, 1995
- Z. Handley, M-J. Hamel, "Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning," Language Learning & Technology, 9, 3, 99-120, 2005
- A. Black, "Speech synthesis for educational technology," Proc. ISCA Workshop on Speech and Language Technology in Education (SLaTE), 2007
- H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds," Acoustical Science and Technology, 27, 6, 349-353, 2006
- R. Kubo et al., "/r/-/l/ perception training using synthetic speech generated by STRAIGHT algorithm," Proc. Spring Meeting of Acoust. Soc. Japan, 1-8-22, 383-384, 1998 (in Japanese)
- S. Kato, G. Short, N. Minematsu, C. Tsurutani, K. Hirose, "Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing," Proc. Int. Workshop on Speech and Language Technology in Education (SLaTE), 2011
- Y. Tsubota, T. Kawahara, and M. Dantsuji, “An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors,” ReCALL Journal, 16, 1, 173-188, 2004
- Y. Tsubota, T. Kawahara, and M. Dantsuji, “Recognition and verification of English by Japanese students for computer-assisted language learning system,” Proc. ICSLP, 1205-1208, 2002
- K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji. “Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system,” Proc. ICSLP, 749-752, 2002
- A. Raux and T. Kawahara, “Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning,” Proc. ICSLP, 737-740, 2002

References

- H. Wang, C. J. Waple, and T. Kawahara, “Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition,” Speech Communication, 51, 10, 995-1005, 2009
- M. Eskenazi, A. Kennedy, C. Ketchum, R. Olszewski, and G. Pelton, “The NativeAccent pronunciation tutor: measuring success in the real world”, Proc. SIG-SLaTE 2007.
- H. Meng, W-K. Lo, A. M. Harrison, P. Lee, K-H. Wong, W-K. Leung and F. Meng, “Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience,” Proc. APSIPA, 2010.
- D. Luo, N, Minematsu, Y. Yamauchi, K. Hirose, “Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences”, Proc. SLaTE, 2009
- Ehsani, F., Bernstein, J., Najmi, A., Todic, O., “Subarashii: Japanese interactive spoken language education,” Proc. Eurospeech, 681-684, 1997
- Zechner, K., Higgins, D., Xi, X., “SpeechRater: a construct-driven approach to scoring spontaneous non-native speech,” Proc. SLaTE, 2007
- S. Lee, H. Noh, J. Lee, K. Lee and G. G. Lee. “POSTECH Approaches for Dialog-Based English Conversation Tutoring,” Proc. APSIPA, 2010
- M. Raab, R. Gruhn and E. Noeth, “Non-native speech databases,” Proc. ASRU, 2007
- N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. Int. Conf. Acoustics (ICA), 557-560, 2004
- N. Minematsu, K. Okabe, K. Ogaki, K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," Proc. INTERSPEECH, 1481-1484, 2011
- N. Minematsu, C. Guo, and K. Hirose, "CART-based factor analysis of intelligibility reduction in Japanese English," Proc. EUROSPEECH, 2069-2072, 2003

Japanese CALL system: CALLJ @Kyoto Univ. [Wang 2009]

- Exercise of basic sentence production (text & speech), given a image scene
- Key features
  - Dynamic generation of questions & ASR grammar network with error prediction
  - Interactive hints



H.Wang, C.J.Waple, and T.Kawahara.
Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition.
Speech Communication, Vol.51, No.10, pp.995--1005, 2009.

Japanese CALL system: CALLJ @Kyoto Univ. How to Try

- Windows only.

0. (Unzip CALLJ1.5.zip).
1. Move to the directory **CALLJ**.
2. Click "**StartCALLJ**".
3. Create your account by clicking "**New**" in login window for the first time.

- You need some knowledge on Japanese.