Prosody Generation

Advanced Signal Processing Seminar SS 2010

Markus Froehle

Signal Processing and Speech Communication Laboratory

June 25, 2010

Outline

Introduction

Text-to-speech synthesis (TTS) systems have to generate speech from text which is:

- natural (i.e., sounding like a human)
- meaningful (i.e., sounding like a human who understands the contents of the text)

Introduction

Text-to-speech synthesis (TTS) systems have to generate speech from text which is:

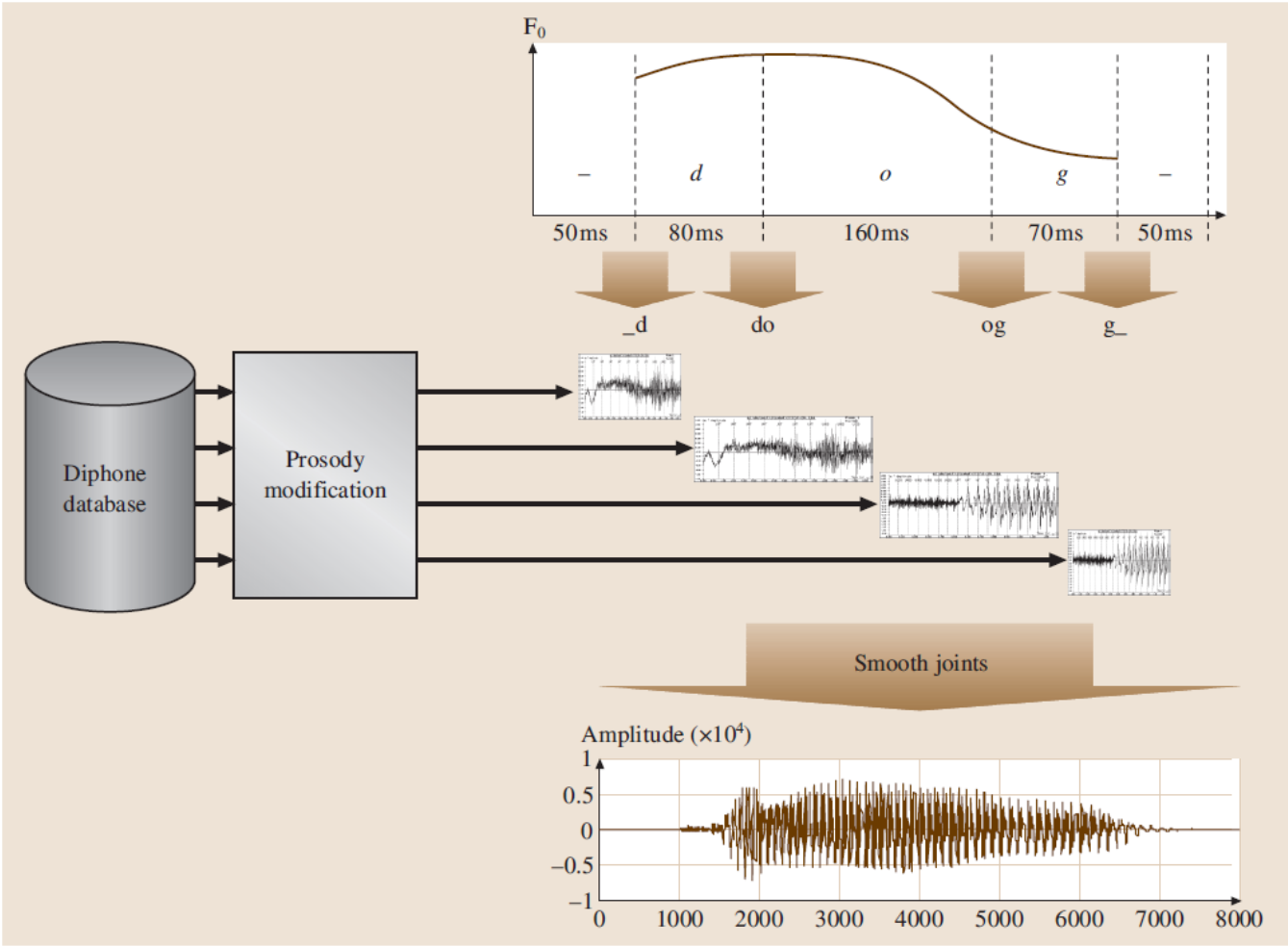
- natural (i.e., sounding like a human)
- meaningful (i.e., sounding like a human who understands the contents of the text)

Several steps are involved to generate speech from text:

- analyse text input
- split single words into syllables
- assign these to specific phoneme sequences
- determine a specification of the prosody
- generate speech with all these extracted features

Example

Diphone Based Text-To-Speech Synthesis



Outline

Prosody

The term prosody is generally used to refer to aspects of a sentence's pronunciation which aren't described by the sequence of phones derived from the lexicon [5].

- pitch
- (sentence) melody
- (speech) rhythm
- loudness

Prosody

The term prosody is generally used to refer to aspects of a sentence's pronunciation which aren't described by the sequence of phones derived from the lexicon [5].

- pitch
- (sentence) melody
- (speech) rhythm
- loudness

Acoustic parameters:

- fundamental frequency F_0 (correlates with pitch and melody)
- segmented duration (correlates with rhythm)
- signal intensity (loudness)

Intonation is the rise and fall of the fundamental frequency F_0 of the voice in speech.

Linguistic Factors of Speech Prosody

... are language specific

Speech prosody used for:

- Distinguishing different meanings of a word:
  - type of F_0 movement within certain syllables
  - different phone duration patterns
- Semantic structuring of utterances
  - Prosodic phrasing: group words together
  - indicate relationships between phrases

  Example:
  - "Charles the first king of England"
    "(Charles the first) king of England"
    "(Charles) (the first king of England)"

Linguistic Factors of Speech Prosody

- Emphasizing of words
  - focus
  - intensity

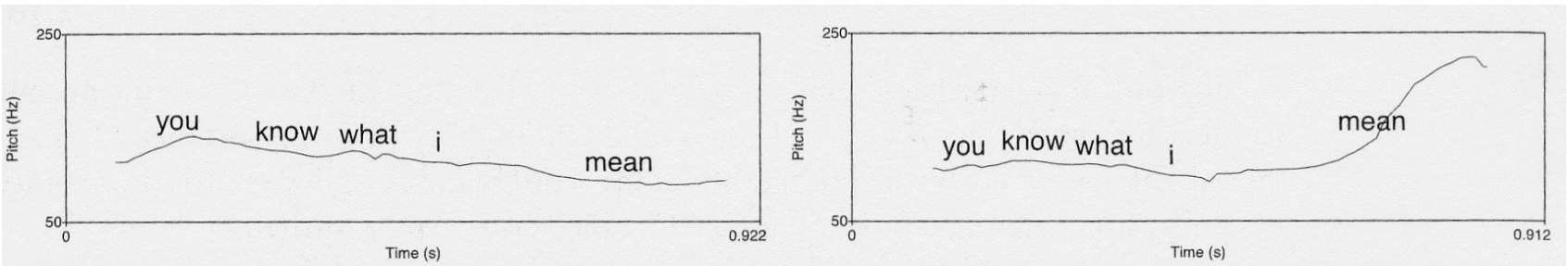
  done by saying it:
  - louder
  - slower
  - varying F_0 during the word
  - making it higher

  Example:
  "I'm a little SUR**PRISED** to hear it **CHAR**ACTERIZED as UP**BEAT**."

Linguistic Factors of Speech Prosody

- Indication of sentence modality
  - statement
  - exclamation
  - question (yes/no)
  - partial question (wh-)
  - parenthesis

Example:



Outline

Representations of Prosodic events

- abstract phonological descriptions of prosodic events
- qualitative phonetic descriptions of alignment and type of prosodic events
  - Pierrehumbert's intonation model (1980)
  - ToBI labelling system (1991, 1994)
- quantitative phonetic descriptions of time alignment and size of prosodic events

Pierhumberts intonation model (1980)

- rule based system for intonation modelling
- considers intonation to be a sequence of high (H) and low (L) tones

H and L tones are the building blocks of larger tone units:

- pitch accents
- phrase accents
- boundary tones

ToBI labelling system

ToBI = *tones and break indices*

- developed on research meetings in 1991 and 1994
- based on Pierrhumbert's theory of intonation
- standard for describing American English intonation

- has been transcribed to other languages and dialects

ToBI labels can be used to generate F_0 contours

- Black and Hunt's approach (1996)

ToBI labelling system cont'd

A ToBI labelling for an utterance consists of three tiers each related (through time) to a speech waveform:

- Labels: pitch accents, phrase accents, boundary tones
- Break indices: one of four levels of prosodic breaks
- Miscellaneous: background noise, coughing, laughing, dis-fluencies, etc.

Show: ToBI labelled sound file

Outline

Generation Methods

- Knowledge-based or Rule-based Methods
  ... historic, knowledge of linguistic experts needed
  - Klatt duration model (1973)
  - Phonetic realization rules for Pierrehumbert's intonation model (1981)
- Data-based, Concatenative Methods
  ... apply the prosody of stored, natural speech units to synthesize the prosodic contours of new utterances
  ... close-domain applications only
  - unit-selection TTS synthesis systems

Generation Methods cont'd

- Data-based, Statistical Methods
  ... apply machine learning techniques to estimate unknown parameters
  - parametric estimation methods
    ... assume model and fit model to the data set
    - Black and Hunt's model (1996)
    - Sums-of-products approach
  - non-parametric estimation methods
    ... determine model entirely through data
    - classification and regression trees (CARTs)
    - artificial neural networks (ANNs)
    - hidden Markov models (HMMs)

The Klatt Duration Model

Rule-based approach for duration modelling developed by Klatt (1973)
Uses rules to model how the average or 'context-neutral' duration of a phone d is lengthened or shortened by context, while staying above a minimum duration d_{min}
It assums that:

- each phonetic segment type has an inherent duration specified as one of its distinctive features: $d_{inherited}$
- each rule results in a percentage p increase or decrease in the duration of the segment, but
- the segment cannot be compressed shorter than a certain minimum duration: d_{min}

The Klatt Duration Model cont'd

The duration of a phone is then

$$d = \frac{(d_{inherited} - d_{min}) \cdot p}{100}$$

Problem of this approach:

- parameter values are based on small-scale case studies and may be inaccurate

Using "as-is" Prosody in Unit Selection Synthesis

Data-based, Concatenative Method

- uses large corpora of natural speech for training and for concatenative synthesis
  - neutral news-readings recordings
- corpus contains several tokens with different phonetic and prosodic context characteristics
- intonation of the concatenated unit is not modified at all

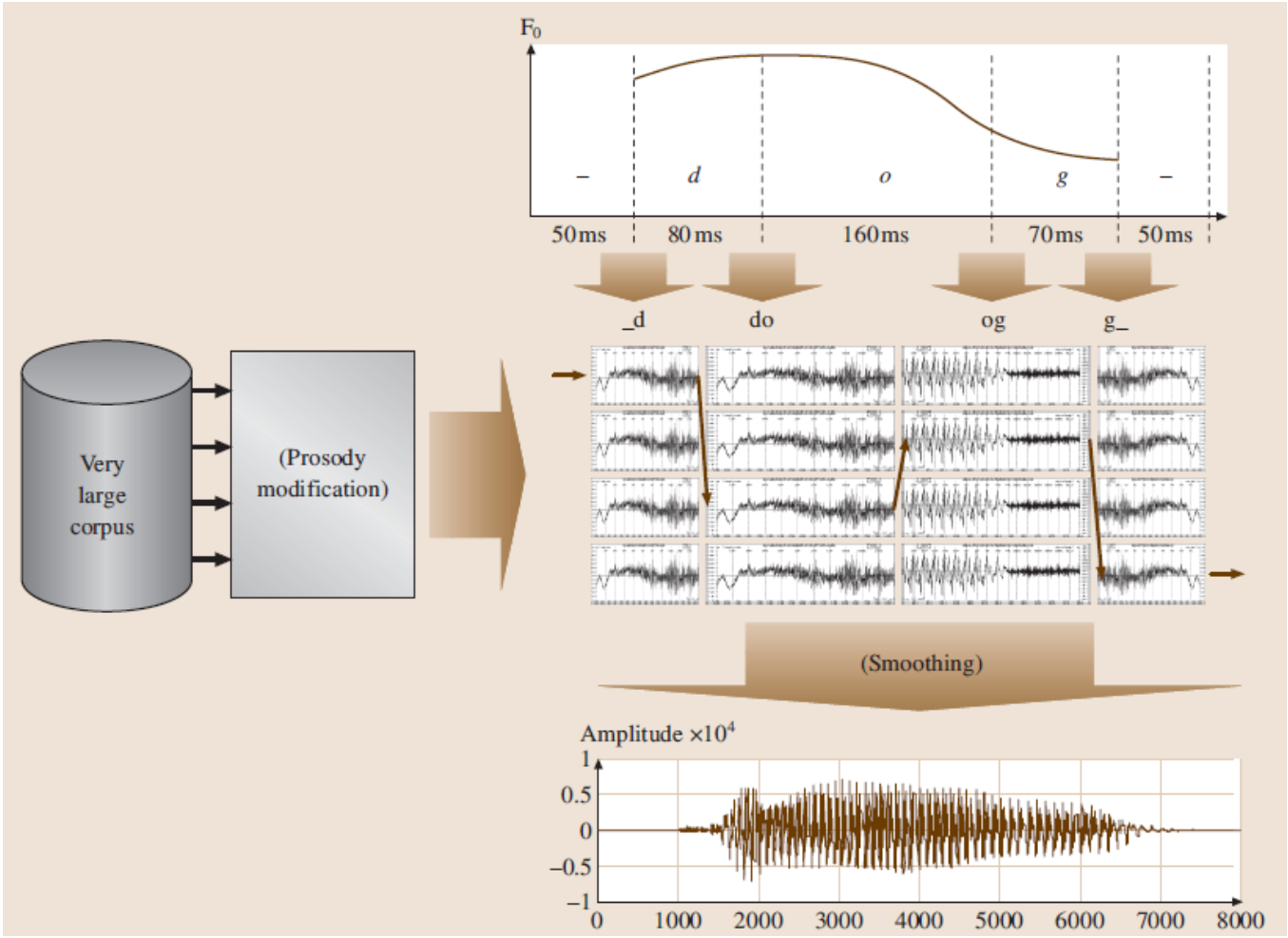
Advantage:

- very natural voice quality

Drawbacks:

- discontinuities, because of unavoidable data sparsity
- not very expressive

Using "as-is" Prosody in Unit Selection Synthesis cont'd



... is not restricted to use diphone units

Generating F_0 contours from ToBI labels using linear regression

Data-based, statistical parametric estimation method developed by Black and Hunt (1996)

Idea:

- predict three F_0 target values (at begin, mid-vowel and end) for every syllable

Prediction formula:

$$target = I + w_1 f_1 + w_2 f_2 + \ldots + w_n f_n$$

f_i ... features
w_i ... weights
I ... constant
weights w_i and I are estimated from data using linear regression

Features f_i

Consist of:

- accents: 5 binary features
- phrase accents and boundary tones: 6 binary features
- break indices: 4 binary features

Algorithm:

- use the same set of features for each syllable
- build three linear regression models: start F_0, mid-vowel F_0 and end F_0
- smooth predicted targets and interpolate for the final F_0 waveform

Experiment and Results

- 45 minutes of an American female speaker reading news
- 14778 syllables: 12000 used for training, rest 2778 for testing
- hand labelled ToBI labels

Baseline model: rule-based data-driven APL method

Results

RMS error and correlation of start, mid-vowel and end F_0:

	Train		Test	
	RMS	Corr	RMS	Corr
start	27.1	0.53	27.4	0.55
mid-v	26.2	0.66	26.1	0.68
end	27.7	0.56	28.4	0.55

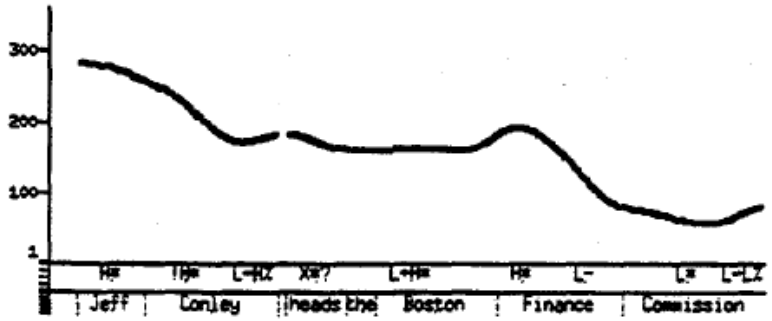
compared with the APL method:

	RMS	Corr
APL	44.7	0.40
LR	34.8	0.62

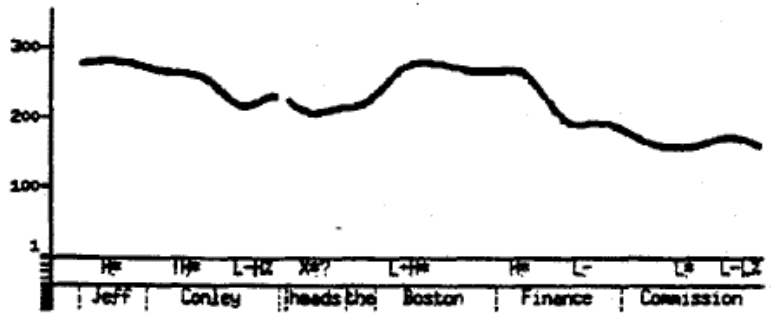
Computed with a $10ms$ frame-by-frame comparison between original and predicted F_0 contour.

Results cont'd

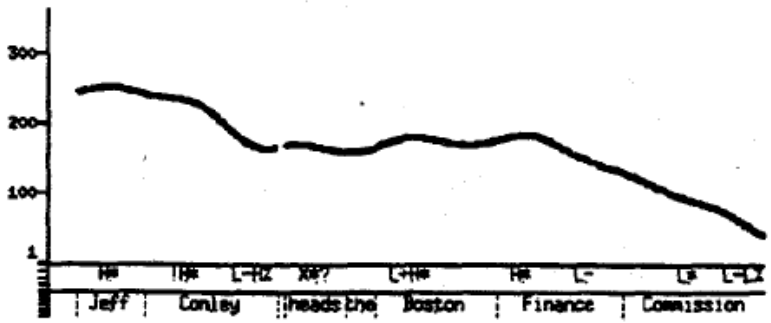
The following graph shows the original smoothed F_0 [SOUND A803S01.WAV]



The contour generated by APL method for the same utterance shows a much more varied contour [SOUND A803S02.WAV]



While the LR method produces [SOUND A803S03.WAV]



Pros and Cons

Pros:

- easy to implement
- if ToBI labels are not available for training the results from a different speaker with the same dialect can be used

Cons:

- there is no way that this technique learn contours not in the training set, or from labels with only a few examples
- the produced F_0 contours were less varied than those generated by APL

Phone duration modelling using gradient tree boosting

Data-based, statistical non-parametric estimation method developed by Yamagishi et al. (2007)

Classification and Regression Trees (CARTs) are

- tree-based non-linear regression algorithms
- produce a binary tree from labelled training data of phonemes based on an optimization criteria

Idea:

- use Gradient Tree Boosting (GTB) instead of conventional approach using regression trees

GTB Algorithm

... gradient tree boosting (GTB) developed by Friedman (2001, 2002)
Define:

- $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ ... explanatory variables
- y ... target value
- $\{y_i, \mathbf{x}_i\}_1^N$ ... a set of training data including N
- GTB algorithm iteratively constructs M different regression trees $h(\mathbf{x}, \mathbf{a}_1), \ldots, h(\mathbf{x}, \mathbf{a}_M)$ from the set of training data and constructs the following additive function:

$$F(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m h(\mathbf{x}, \mathbf{a}_m)$$

GTB Algorithm cont'd

$$F(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m h(\mathbf{x}, \mathbf{a}_m)$$

- β_0 ... initial weight
- β_m ... weight
- $\mathbf{a}_m$ ... vector of parameters for m-th regression tree $h(\mathbf{x}, \mathbf{a}_m)$
- β_m and $\mathbf{a}_m$ are iteratively determined so that a loss function $\Psi(y, F(\mathbf{x}))$ is minimized

Define an additive function that is combined from the first to the m-th regression tree:

$$(\beta_m, \mathbf{a}_m) = argmin_{\beta,\mathbf{a}} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i, \mathbf{a}))$$

GTB Algorithm cont'd

$$(\beta_m, \mathbf{a}_m) = argmin_{\beta,\mathbf{a}} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i, \mathbf{a}))$$

with initial value:

$$F_0(\mathbf{x}) = \beta_0 = argmin_{\beta} \sum_{i=1}^{N} \Psi(y_i, \beta)$$

- equations not straight forward to solve
- Therefore, GTB separately and approximately estimates $(\beta_m, \mathbf{a}_m)$
- used least-square loss function: $\Psi(y, F) = \frac{(y-F)^2}{2}$

Experiment

Languages:

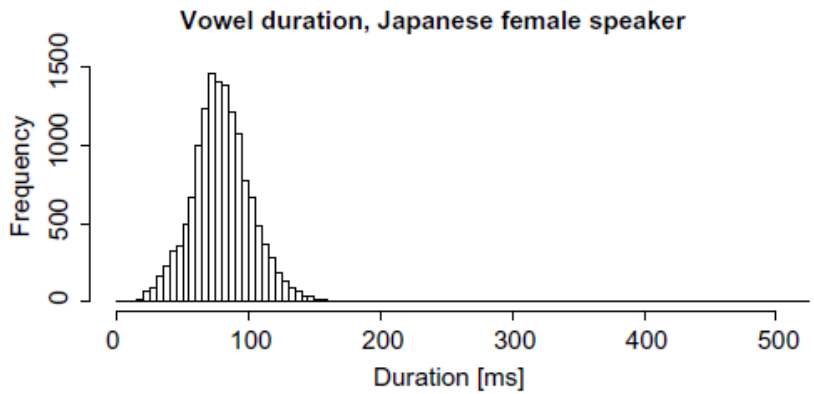
- Japanese
- Mandarin
- English

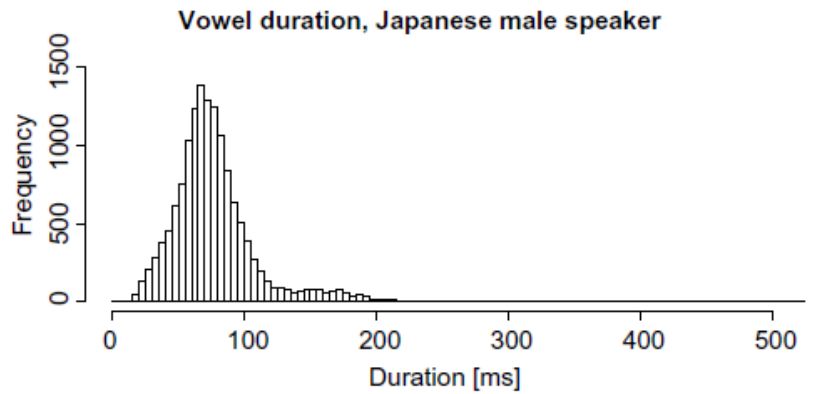
manually labelled phone duration and several explanatory variables of the utterances:

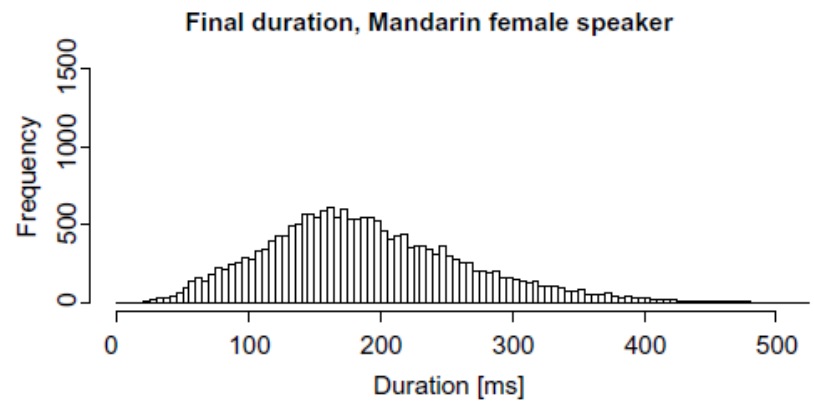
- 53 English explanatory variables:

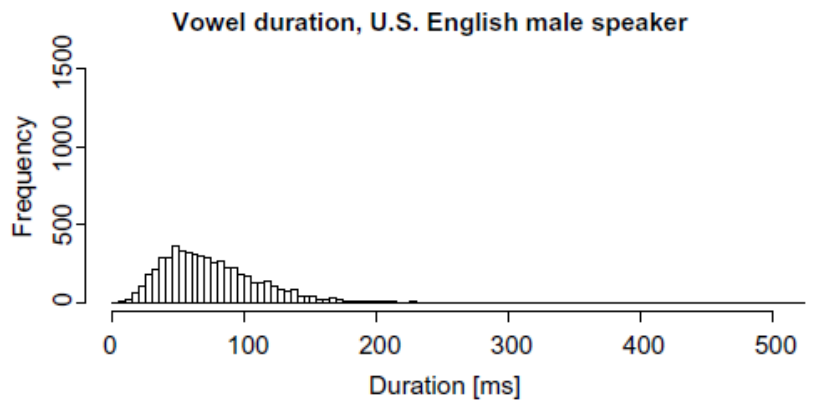
  5 phonetic features, 2 segment-level features, 22 syllable-level features, 12 word-level features, 9 phrase-level features, 3 utterance-level features
- 47 for Japanese
- 58 for Mandarin

Experiment cont'd









Evaluation

- using 5-fold-cross validation on speech databases
- two measures for objectiveness:
  - pseudo R-squared:

$$R^2 = 1 - \frac{\sum_{i=1}^{T} (F(\mathbf{x}_i) - y_i)^2}{\sum_{i=1}^{T} (y_i - \bar{y})^2}$$

  - root mean square error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T} (F(\mathbf{x}_i) - y_i)^2}{T}}$$

- Baseline model: conventional approach using a regression tree

Results

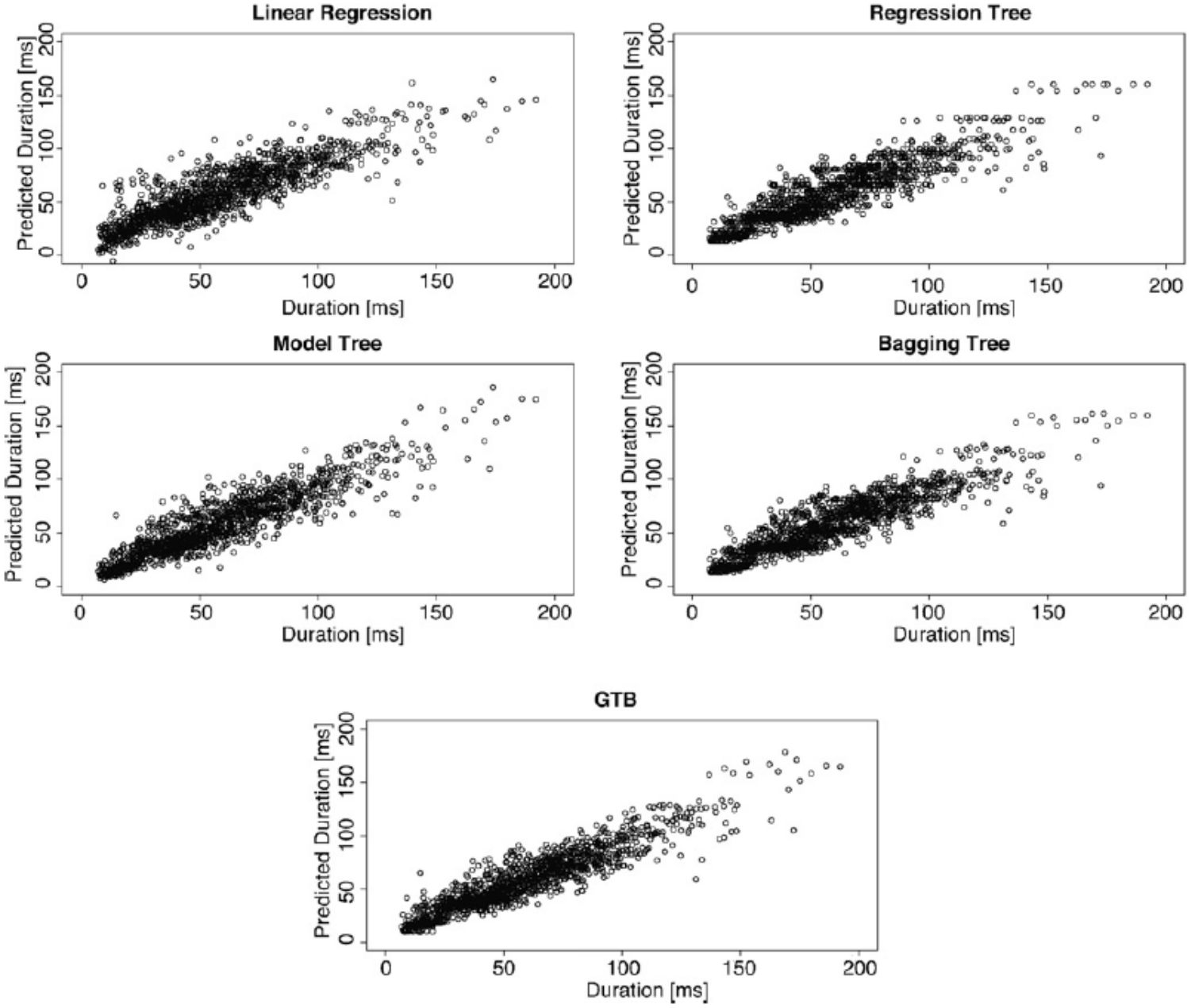
Comparison results with other duration modeling techniques for the Japanese female speaker

Model	R^2	RMSE (ms)
(a) Vowel		
Linear regression	0.58	14.32
Regression tree	0.55	14.94
Model tree	0.58	14.28
Bagging tree	0.57	14.51
GTB	0.61	13.87
(b) Consonant		
Linear regression	0.76	15.37
Regression tree	0.80	13.85
Model tree	0.83	12.94
Bagging tree	0.82	13.47
GTB	0.85	12.08

Results cont'd

- Similar results for Mandarin and English
- also better than a Neural Network approach:

Model	R^2	RMSE (ms)
(a) Vowel		
MLP	0.54	22.65
GTB	0.73	16.12
(b) Consonant		
MLP	0.55	20.22
GTB	0.78	12.77


Linear Regression
Predicted Duration [ms]
Duration [ms]
0
50
100
150
200
Regression Tree
Predicted Duration [ms]
Duration [ms]
Model Tree
Predicted Duration [ms]
Duration [ms]
Bagging Tree
Predicted Duration [ms]
Duration [ms]
GTB
Predicted Duration [ms]
Duration [ms]

Outline

Conclusion

- many different approaches for TTS systems exist
- none of them is perfect

- a good approach is unit-selection, BUT needs a huge database
- Future: Hybrid approach?

References

Black, A. W. and Hunt, A.,
"Generating F_0 Contours from the ToBI Labels Using Linear Regression",
in *Proceedings of ICSLP'96*, 1996, pp. 1385–1388.

Yamagishi, J. and Kawai, H. and Kobayashi, T.,
"Phone Duration Modeling Using Gradient Tree Boosting",
Speech Communication, vol. 50, no. 5, pp. 405–415, May 2008.

Dutoit, T.,
"Corpus-Based Speech Synthesis",
in *Handbook of Speech Processing*, Benesty, J. and Sondhi, M. and Huang, Y. Berlin Heidelberg: Springer, 2008, ch. 21, pp. 437–455.

van Santen, J. and Mishra, T. and Klabbers, E.,
'Prosodic Processing",
in *Handbook of Speech Processing*, Benesty, J. and Sondhi, M. and Huang, Y. Berlin Heidelberg: Springer, 2008, ch. 23, pp. 471–487.

D. Jurafsky and J. H. Martin,
'Speech and Language Processing",
second edition ed. Pearson, 2009.

References cont'd

H. Romsdorfer,

'Polyglot text-to-speech synthesis. text analysis and prosody control",
Ph.D. dissertation, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January 2009.

P. Taylor,

'Text-to-Speech Synthesis",
2009.