



Foot Structure and Pitch Contour Paper Review

-----*-----*-----*-----*-----*-----*-----*

Language Technologies Institute

Carnegie Mellon University

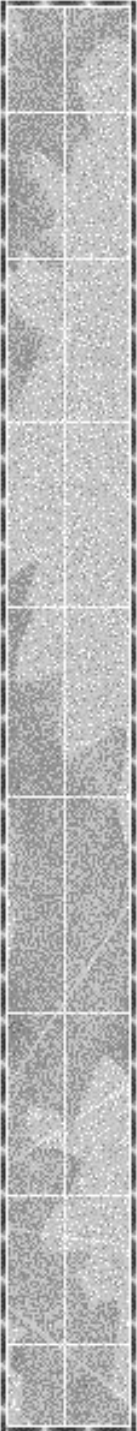
7/22/2004

-----*-----*-----*-----*-----*-----*



Papers

-
- ✱ Esther Klabbers, Jan van Santen and Johan Wouters, “Prosodic Factors for Predicting Local Pitch Shape,” IEEE 2002 Workshop on Speech Synthesis
 - ✱ Esther Klabbers and Jan P. H. van Santen, “Control and prediction of the impact of pitch modification on synthetic speech quality,” Eurospeech 2003
 - ✱ Esther Klabbers and Jan P. H. van Santen, “Clustering of foot-based pitch contours in expressive speech,” SSW5, 2004.



1st Paper: IEEE 2002 Workshop

-
- * Investigate predictive power of different prosodic factoring schemes.
 - * Extend diphone voice by making additional recordings under different prosodic contexts.
 - * Use foot structure to guide choice of prosodic contexts.



Introduction

✱ Problem: corpora typically have 1 example per diphone coming from stressed context

- These examples are sometimes bad matches for prosodic context, and much signal modification (with potential quality degradation) can be necessary.
- Adding many examples to cover more possibilities could lead to a large database
 - Difficult to use in embedded devices
 - Difficult to keep speaker consistent across more examples
 - Need to find good “selection criteria”



Feet and Pitch

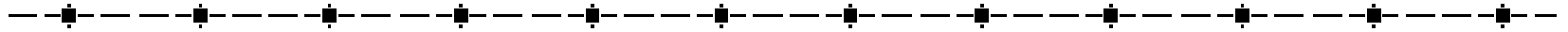
* “Left-headed foot”

- ◆ Sequence of 1 or more syllables, 1st is accented
- ◆ Followed by accented syllable or phrase boundary

* Typical accent, up-down pitch movement

- ◆ Monosyllabic: rise-fall on single syllable
- ◆ Polysyllabic: rise on first, fall on rest

Factorization Schemes



	Simple	Foot	Complex1	Complex2
	Stress {0,1}	Last accent {0,1,2}	Accent{0,1}	Accent {0,1}
	Accent {0,1}	Next accent {0,1,(2)}	Last accent {0,1,2}	Last accent {0,1,2,3}
	Phrase-fin. Syll. {0,1,2 }	Phrase-fin. Foot{0,1,2}	Next accent {0,1,2}	Next accent {0,1,2,3}
Levels	12	19	54	96

Experiments

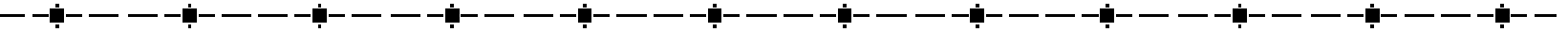
✧ Corpus

- ◆ 472 sentences spoken by a female
- ◆ Segmented and annotated by hand
- ◆ 1493 of 8860 syllables were used
 - Only ones starting with a sonorant

✧ Measures

- ◆ RMSE between one contour and another contour estimated from the second
- ◆ Delta distance

Results



Mean	Simple	Foot	Complex1	Complex2
Levels	12	19	30	48
RMSE	13.1	12.8	12.7	11.9
Delta Distance	11.9	10.9	11.3	10.4



Discussion

-
- ✱ Foot scheme performs better than Simple and similar/better than Complex1
 - ✱ Complex2 performs best but has too many factors.
 - ✱ “Hypothesis 1: The distinction between medial, phrase-final and utterance-final feet is important for predicting pitch contour shapes.”
 - ✱ “Hypothesis 2: The position of the previous accented syllable is irrelevant if the current syllable is the head of the foot.”



Text Corpus Analysis

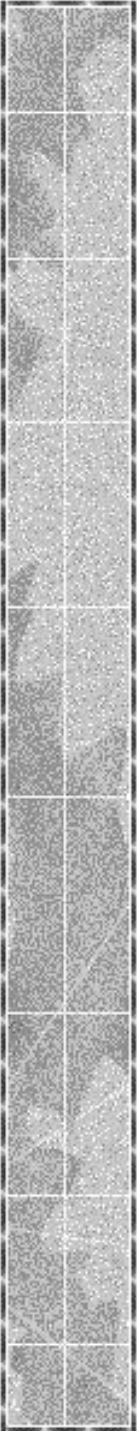
✱ Analyzed large text corpus

- ◆ 359,276 sentences from newspapers, novels, and bible
- ◆ Used Festival to compute foot factor levels for each diphone: 16,926,727 total of 22,865 types
- ◆ Simplified by disregarding consonant position and only having single versions of consonant-consonant diphones: 9,367,407 tokens of 21,458 types
- ◆ Using a standard database of 3353 diphones, only 6020 had to be added to cover 95% of diphone-foot tags.



2nd Paper: Eurospeech 2003

-
- ✱ Continues in the vein of trying to reduce the amount of signal modification necessary by using foot structure to improve selection.
 - ✱ Perceptual experiment to investigate degradation caused by pitch modification
 - ✱ Correlation of weighted perceptual score with different pitch and delta pitch distances



Speech Corpus Analysis

✦ Same prosodic factorization as 1st paper

✦ Corpora

- ◆ Duration corpus: corpus from 1st paper

- ◆ Foot Corpus I

- Recorded to testing effect of position on pitch contour
- 285 sentences, spoken by a highly-expressive female
- Each sentence target is an all-sonorant CVC syllable

- ◆ Foot Corpus II:

- Instructed speaker to be less expressive, speaker uncomfortable

Distance Measures

✱ Tried various distance measures

$$D_p = \sum (\log_{10}(F_{0i}) - \log_{10}(F_{0j}))^2$$

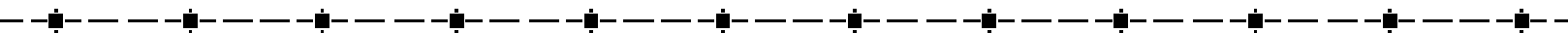
$$D_{wp} = \frac{\sum E(\log_{10}(F_{0i}) - \log_{10}(F_{0j}))^2}{\sum E}$$

$$D_{\Delta p} = \sum (\Delta \log_{10}(F_{0i}) - \Delta \log_{10}(F_{0j}))^2$$

$$D_{w\Delta p} = \sum E(\Delta \log_{10}(F_{0i}) - \Delta \log_{10}(F_{0j}))^2$$

$$\text{where } E = \sqrt{E_i \times E_j}$$

Results



- ✦ Foot annotation scheme performed better than Simple for all 3 corpora and was generally better than complex
- ✦ It appeared that some levels in the Foot scheme could be collapsed further
 - ◆ For Head, Doesn't matter whether unstressed syllables follow
 - ◆ For unstressed syllables, only matters whether they are immediately preceded by the head
 - ◆ For all syllables, important if foot is phrase-medial, phrase-final with continuation rise, or utterance-final.
 - ◆ New 12-level factorization scheme that is still better than simple



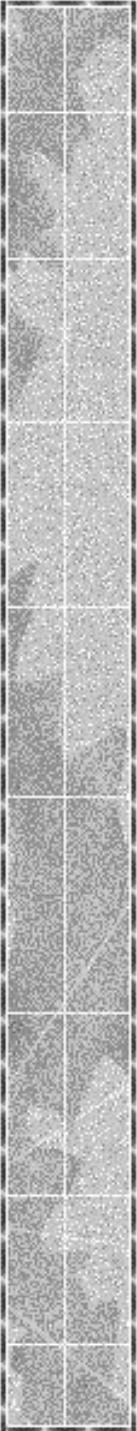
Perceptual Experiment

- ✧ Use OGiresLPC algorithm
- ✧ Use data from Foot Corpus I
- ✧ Sentences had carrier phrase and target word
 - ◆ Target word was sonorant CVC syllable from corpus
 - ◆ Two versions: one where syllable is in same prosodic context, another where it is in different context
 - ◆ Sentence parts were concatenated with Snack
 - 20ms pause inserted between carrier phrase and target word
- ✧ Participants compared pairs on 7-point scale



Results

-
- ✦ Computed weighted score for each sentence, based on z-score normalization
 - ✦ Used linear regression with different distances to try to predict weighted scores
 - ✦ At first, appeared that pitch distance and delta distance caused highest variance
 - ✦ With more varied material, weighted distances might give better correlations.
 - ✦ Direction of pitch change important.
 - ◆ 2 new distance measures were created
 - ◆ Decreasing pitch was worse than increasing pitch



3rd Paper: 5th ISCA SSW

-
- ✱ Concerned with categorizing foot-based pitch contours in expressive speech
 - ◆ Clustering instead of prediction
 - ✱ Classifying emotions in speech is problematic, so focusing on what pitch contours actually occur



Models

* TTS system used Generalized Linear Alignment Model

- ◆ Pitch contour consists of phrase curves, accent curves, and segmental perturbation curves
- ◆ Phrase curve has two linear components
 - Phrase start to syllable with nuclear pitch accent
 - There to end, with steeper decline

* This paper uses Simplified Linear Alignment Model

- ◆ Assumes accent is realized by up-down movement, where location depends on # syllables in foot



Corpus

-
- * 2 children's stories by Beatrix Potter
 - * Read by semi-professional female speaker
 - * 10 minutes of speech, not counting pauses
 - * 2929 syllables
 - * 128 sentences



Annotation

-
- * Automatic phoneme segmentation by CSLU's phonetic alignment system
 - * Phonetic transcription from Festival
 - * Phonemes checked and alignments hand-corrected with Wavesurfer
 - * Syllable transcription created by hand and aligned with phoneme labels
 - * ESPS get_f0 used to extract pitch every 5ms
 - Wrote Wavesurfer plug-in to interpolate with lines



Pitch Normalization

◆ Pitch contours are different lengths and need to be normalized for comparison.

- ◆ Simple interpolation doesn't work because peak location tends to differ between monosyllabic and polysyllabic feet.
- ◆ Predicted peak locations were used to split intervals, and 50 points were sampled on each side.



Analysis

✱ Distances between pitch contours were calculated as:

✱ $1 - \text{cor}(F_{0i}, F_{0j})$



Clustering

-
- ✱ Used S-PLUS “hclust” method for clustering (non-metrical hierarchical)
 - ◆ Each object gets own clustered, then clusters joined until only 1
 - ◆ Used “ward” method: minimum variance method that finds compact spherical clusters
 - ◆ Final number of clusters determined empirically by looking and listening



Results

-
- * 6 clusters were selected
 - * The paper has figures of medians of z-normalized pitch contours for each cluster.
 - * There is a also a table showing bigram relative frequencies, with some discussion.



Conclusion

- ✦ Authors feel this paper has shown assumptions made in Generalized Linear Alignment Model are correct.
- ✦ Discoveries
 - ◆ “two feet (most frequently occurring at the end of a minor or major phrase) can be connected by what seems to be a different type of phrase curve consisting of an increasing movement on the first foot and a decreasing movement on the last foot.”
 - ◆ “continuation rise which was always assumed to be present at minor phrase boundaries was only observed in fewer than 10% of feet occurring at the minor phrase boundary in this corpus.”
- ✦ Need to confirm these discoveries for other speakers