Detecting Rhythmical Prominence in Speech by an Optimized Convolution Kernel

Christina Orphanidou and Greg Kochanski

Phonetics Laboratory, University of Oxford, UK

***P**atterns of prominence define the rhythmicity of speech, an important characteristic of stressed-time languages, like English. We investigated which acoustic properties of the speech signal mark **rhythmical prominence**.*

A production experiment was conducted during which subjects repetitively read out speech to a metronome, trying to match stressed syllables to its beat. We then searched for the function of the speech signal that best predicts the timing of the metronome ticks. We found that the most important factor is the contrast in ***specific loudness*** between a syllable and its neighbours in an approximately 360 millisecond window centered on the syllable in question relative to an approximately 800 millisecond-wide symmetric window.

ANALYSIS

We searched for the function of the speech signal that best predicts the metronome ticks. We applied it to the waveform and got a set of predictions. We computed the accuracy of these predictions compared to the metronome ticks and adjusted the parameters to maximize the accuracy.

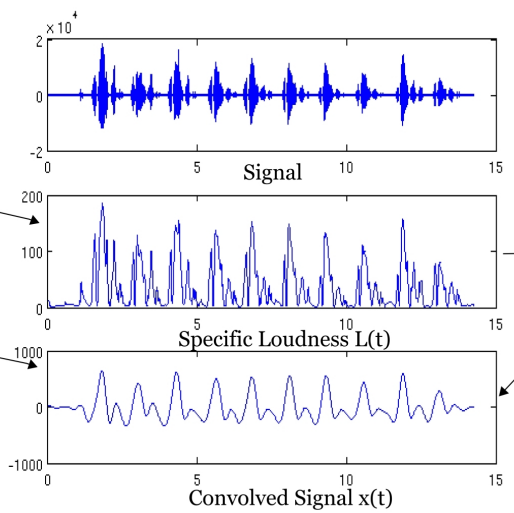
The algorithm

We first compute the ***specific loudness*** *L(t)*

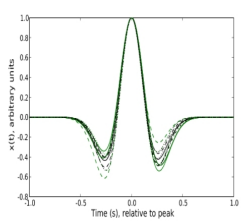
We then convolve it with a kernel *K* to yield *x(t)*.

$$x = K * (L \cdot g)$$

g will include different acoustic properties of the speech signal.



Convolution kernel,K



Optimal samples of the convolution kernel, K. The maxima of the curves are aligned at t=0.

EXPERIMENTAL METHODS

Stimuli

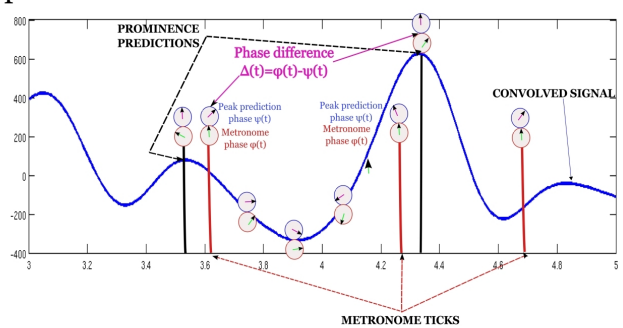
A set of 48 short phrases (4-6 syllables with at least one polysyllabic word) were selected from Project Gutenberg.
Examples: *"Nothing Matters", "Indeed it had", "Talking of wandering", "Probably not".*

Procedure

After some warm-up tasks a metronome was connected to an earphone and subjects read out 10 consecutive repetitions of 24 randomly chosen sentences from the stimuli set.

Optimization

We optimize the acoustic parameters by minimizing the ***phase difference*** Δ(t) between the predictions of prominence and the metronome ticks.



We then compute $I = \frac{1}{L}\int exp(i \cdot \Delta(t)) \cdot dt$

We evaluate |I| for 90000 randomly chosen combinations of parameters and take the one that produces the largest absolute value.

Acoustic Properties

We test different forms of ***g*** based on the following acoustic properties of the speech signal:

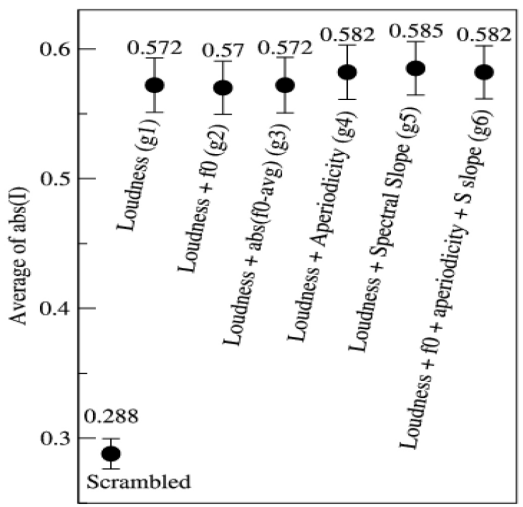
1. Loudness
2. Fundamental frequency
3. Voicing
4. Aperiodicity
5. Spectral slope

We gratefully acknowledge support from the UK's Economic and Social Research Council: RES-000-23-1094

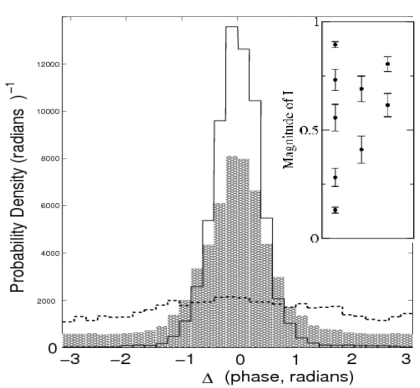
Detecting Rhythmical Prominence in Speech by an Optimized Convolution Kernel

Christina Orphanidou and Greg Kochanski
Phonetics Laboratory, University of Oxford, UK

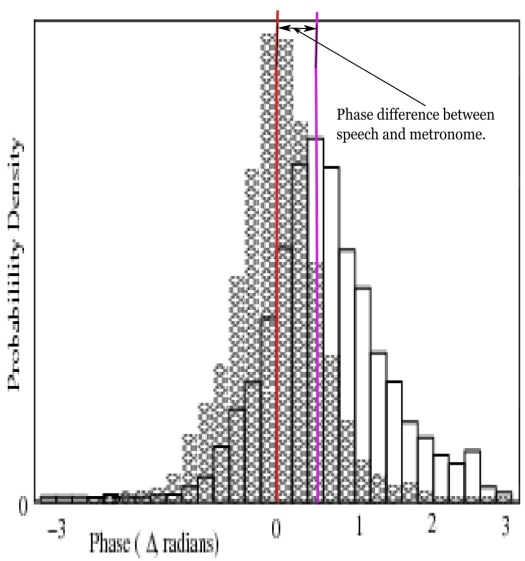
RESULTS



Values of |I| averaged over the corpus. The lower left measurement is the baseline, where acoustic data is shuffled with respect to ticks. Other conditions correspond to different combinations of acoustic parameters.
L is the most important correlate of the ticks whereas other factors add very little improvement to the prediction.



Phase histograms for three different subjects. The subjects shown have the largest (top), median (middle) and smallest (bottom) value of |I|.



Phase histogram (Δ) for a typical speaker (outline) and histogram of Δ relative to the average phase of each utterance (filled). The width of the histograms show timing inconsistensies between the subject's speech and the metronome.

DISCUSSION/CONCLUSION

1. The specific loudness *L* is an important correlate of the ticks while other factors such as *fo*, spectral slope, voicing and aperiodicity add very little improvement to the prediction.

2. Ticks were found to be correlated with the properties of a region larger than a single syllable.

3. Subjects differed substantially in their performance the task may therefore be a useful way of measuring the ability to produce metrical patterns.

4. The results were repeated using only the best performing subjects with no substantial differences observed.

In English, people mark the beat of repetitive speech by speaking loudly relative to the immediate neighbourhood. This neighbourhood is approximately one syllable on either side of the beat, at the speech rates we studied. The critical factor is the average loudness over a 360 millisecond interval. Other factors such as *fo* are not strongly correlated with the beat.