Automatic Time-Alignment of Spoken Sentences

Inseok Heo and William Sethares

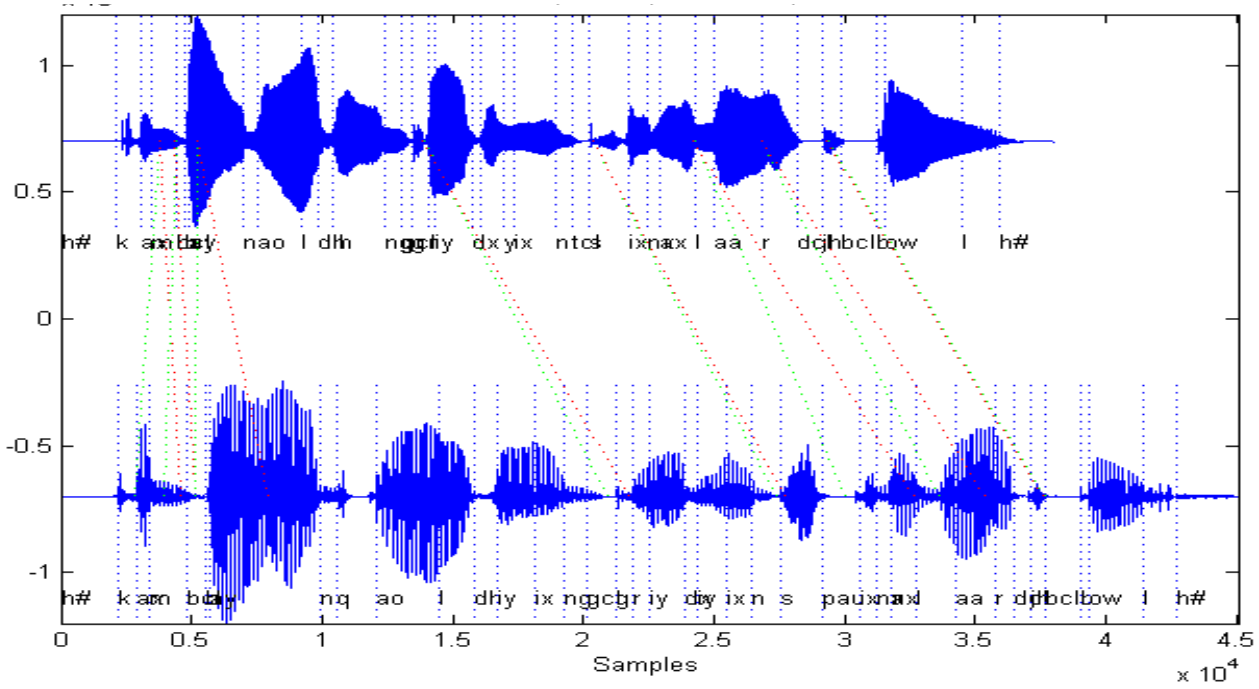
Two spoken sentences are never identical, even if they contain the same textual material. Basic to the study of how speakers differ is the ability to (automatically) parse the sentences and align the relevant utterances, to locate the times when the phonemes of one speaker align with the (same) phonemes of another. This enables an analysis of the micro-timing of events in the speech and has uses in linguistic studies, in speech recognition systems, in speech therapy, and in audio/video synchronization.

The speech accent archive

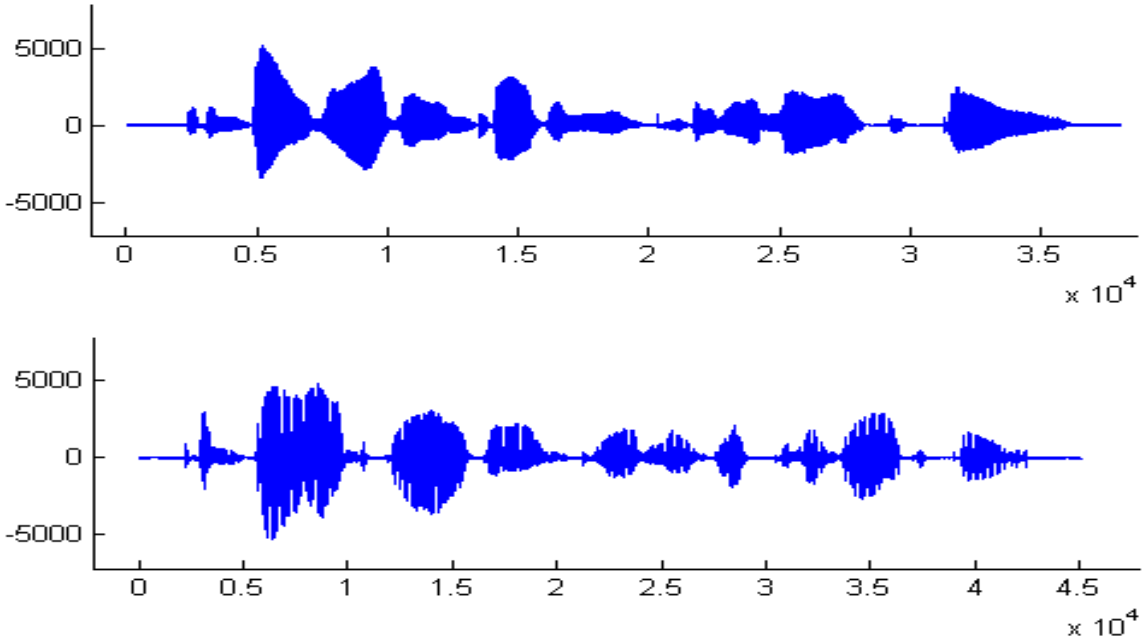
http://accent.gmu.edu

Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

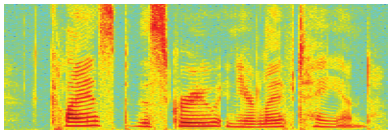
The timing of everyone's speech is different:



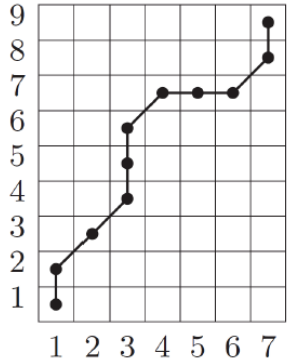
Goal is to align the speakers in time



spectrogram



Feature Extraction



Dynamic Time Warping

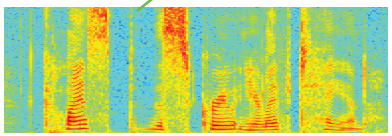
table of aligned time points

time aligned speech

DTW

Variable Time Stretch

Feature Extraction



spectrogram

analysis

synthesis by overlap/add or by phase vocoder

Related Literature

AUTHOR	FEATURE	Win len	DIST	DTW constraints	T-S	SCENARIO	EVALUATION
Chamberlain '83	19-Filter bank Log Power	20ms	L2	1. Asymmetric 2. Symmetric (Sakoe-Chiba)	None	Isolated word recognition	None
Verhelst '97	LPC	30ms	L2	Sakoe-Chiba	WSOLA	Post-recording voice in film	None
Resch '03	MFCC13	20ms	L2	Modified Smoothing	WSOLA	Sentence alignment	Listening test
Ellis '03	Power Spectrum	32ms	COS	Sakoe-Chiba	PV	Sentence alignment (Polyphonic music transcription)	Listening test
Soens '05	MFCC14	30ms	L2	Splitted warping path	WSOLA	Sentence alignment	Listening test
King '12	PLCA (same speaker)	32ms	COS	Sakoe-Chiba	None	Dialogue replacement (noisy environment)	Percentage of correctly aligned frames

There are some details to take care of...

- choice of parameters in spectrogram (FFT-size, hop length, window...)
- choice of features (use FFT? cepstral coefficients? bark scale, zero crossings...)
- distance function in DTW (Euclidean, cosine, weighted norm...)
- how to best do time stretching in resynthesis

We noticed that many mistakes seemed to occur because the system missed the “start” of words. How to emphasize the attacks?

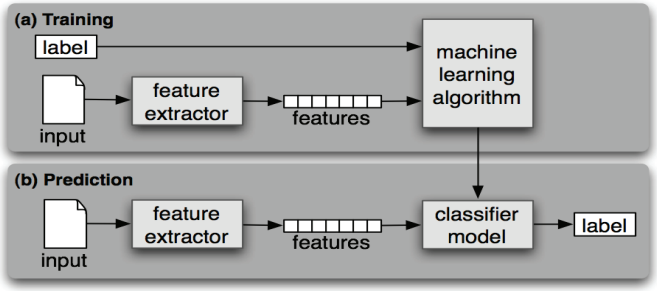
Attacks are short... vowels are long...

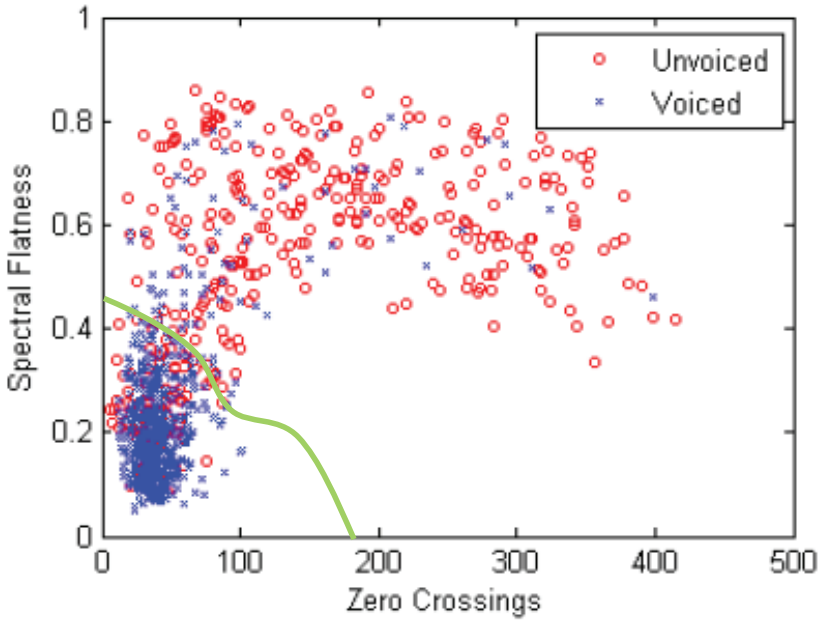
What if we could use more/smaller windows on the attack portions of the words and longer windows on the sustained parts (such as vowels). Then the DTW would have more frames to work with during the alignment process.

Can do two passes through the audio: the first time do a regular spectrogram and classify each frame into sonorant/obstruent (voiced/unvoiced) segments. The second pass uses the classification to choose where to place short windows (at the attacks/obstruent segments) and where to use large windows (on the sonorant/vowel segments). We call this second pass a variable window spectrogram.

First Pass Through the Audio

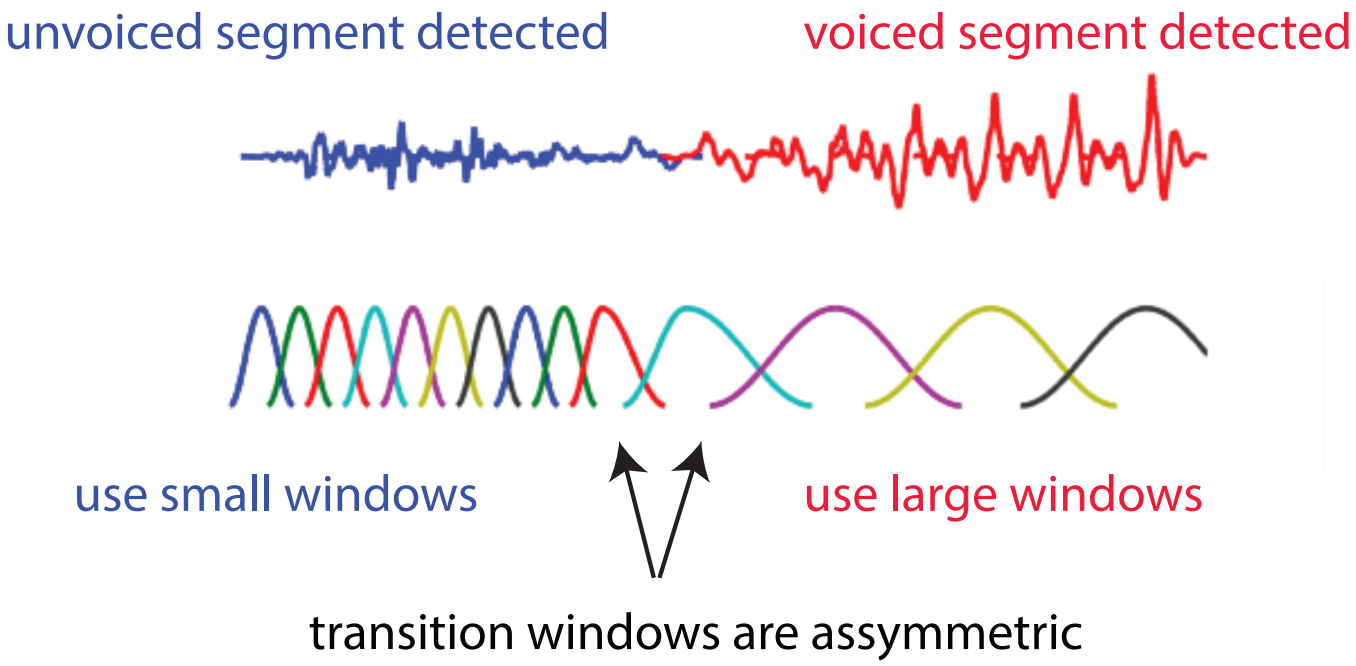
Use a standard (fixed-window) spectrogram to build a classifier (SVM with Gaussian kernel) to distinguish the attack and sustain portions of the speech.





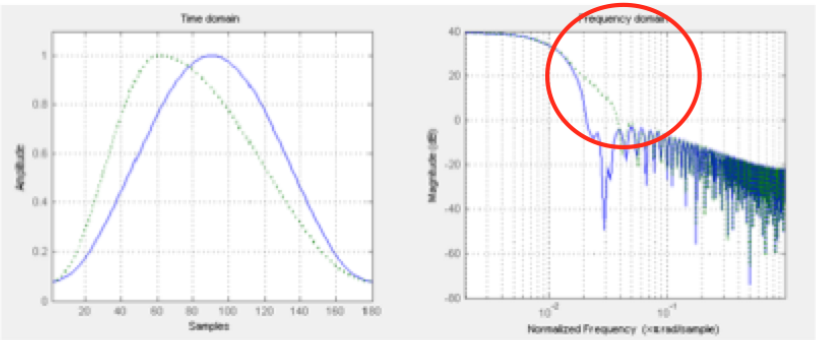
Variable Width Spectrogram

Use the classification in the first pass to specify the windowing for the second pass through the audio.

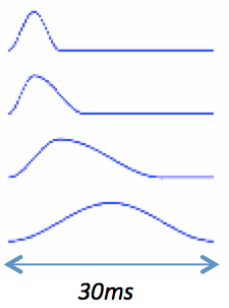


Asymmetric window

Spectral leakage of asymmetric hamming window

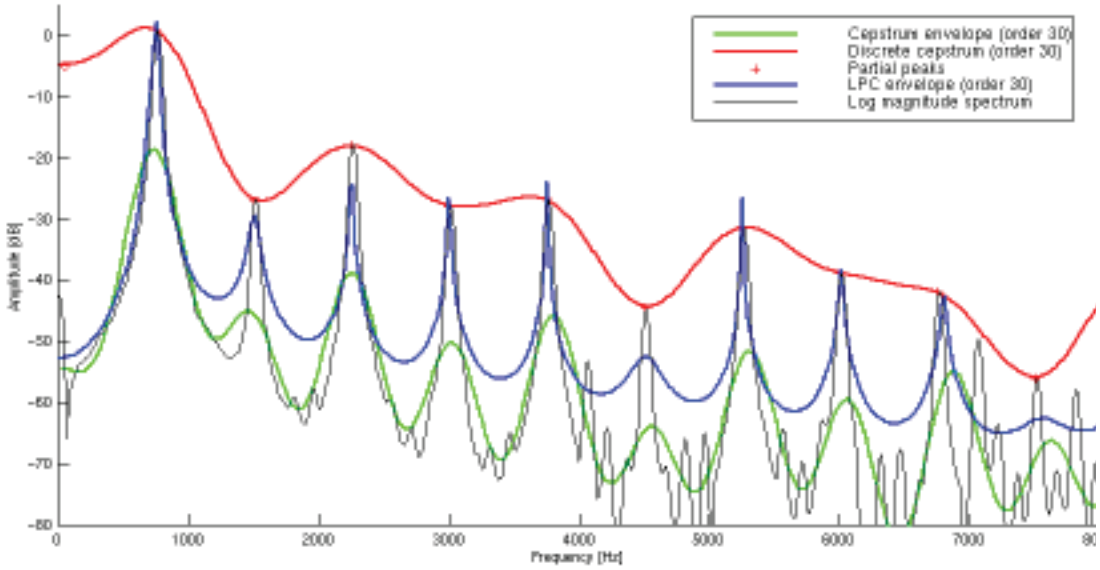


Signal	Type	Window Length	Zero padded	DFT length	frequency resolution
Unvoiced(S)	I	7.5 ms	22.5 ms	30 ms	133 Hz
Transition1(A)	II	11.25 ms	18.75 ms	30 ms	89 Hz
Transition2(A)	III	22.5 ms	7.5 ms	30 ms	44 Hz
Voiced(S)	IV	30.0 ms	0.0 ms	30 ms	33 Hz



Acoustic Features

- LPC (Linear Predictive Coding): All-pole resonance filter, speech codec
- MFCC (Mel-frequency cepstral coefficient)
- PLCA (Probablistic Latent Component Analysis)
- Magnitude of FFT vectors



All Together Now...

Future Work

- Can the origin of an accent be determined from temporal information alone? Use accent data base as training set – take an unknown and see if it can be classified correctly.
- An app to help users learn to speak with timing more like a native speaker
- A variable-window phase vocoder? Merge the overlap/add and PV methods?
- Examine micro-timing of performances: Hamlet's soliloquy, poem, or a chant... emotional effects and/or personal differences.