MEASURING FINAL LENGTHENING FOR SPEAKER-CHANGE PREDICTION

Anna Hjalmarsson & Kornel Laskowski
Speech, Music and Hearing, KTH, Sweden

Goals

- To explore pre-silence final lengthening as a predictor for next-speakership in dialogue
- To evaluate an automatic measure of spectral envelope change, Mel-spectral flux (MSF) by comparing the performance of MSF to a transcription-mediated measure

Corpus

- Human-human dialogue (spontaneous negotiation)
- Segmented into 1897 inter-pausal units (IPU; a sequence of words inter-separated by silences no longer than 100 ms)
- Of 1897 IPUs, 942 (49%) were followed by turn-medial silence and 955 (51%) were followed by the other party speaking

Mel-spectral flux (MSF)

- "Spectral flux" (SF), or "delta spectrum magnitude", but applied to the Mel-spectral envelope

$$MSF = -\text{logit}\left(\frac{\mathbf{m}_L \cdot \mathbf{m}_R}{\sqrt{\mathbf{m}_L \cdot \mathbf{m}_L}\sqrt{\mathbf{m}_R \cdot \mathbf{m}_R}}\right)$$

- $\mathbf{m}_L$ and $\mathbf{m}_R$ are the left and right Mel-spectral envelopes of the left and right portions of each analysis frame
- MSF is evaluated every 8ms, over analysis frames 64ms in duration, and averaged over the last 400ms of each IPU

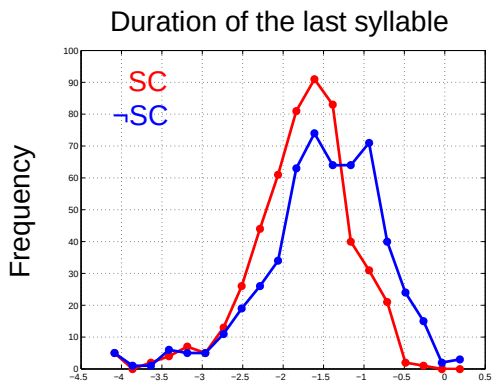
Results

Binary logistic regression was used to estimate the probability that an IPU is SC or ¬. The DEVSET are 10-fold crossvalidation results, while the EVALSET are single-fold results with models trained on all of DEVSET

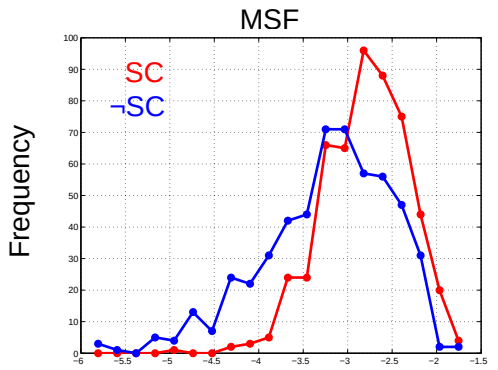
Data	Features	(#)	DevSet Acc	DevSet AUC	EvalSet Acc	EvalSet AUC
	DUR1	(1)	59.0	62.7	62.5	64.7
all	MSF	(1)	65.9	72.2	64.2	69.1
	+DUR1	(2)	67.5	72.5	65.3	70.2
	DUR2	(1)	57.3	59.5	54.3	55.8
	DUR1	(1)	59.0	62.4	62.9	65.0
	+DUR2	(2)	64.3	66.5	62.5	66.3
sub	MSF	(1)	66.2	73.0	64.9	69.7
	+DUR2	(2)	66.4	73.5	66.2	70.4
	+DUR1	(2)	68.0	73.2	65.5	70.6
	+DUR2	(3)	68.4	73.8	67.4	71.4

Data sets (# ¬SC,SC): DevSet(533,517) and EvalSet(409,438)

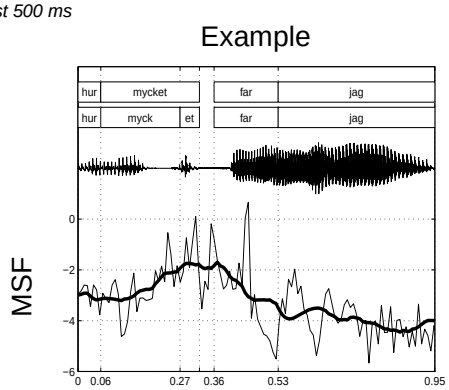
Final-lengthening in SC and ¬SC IPUs



Unnormalized distribution of the natural logarithm of the duration (in seconds) of the last syllable



Unnormalized distribution of the MSF feature averaged over the last 500 ms



Conclusions

- Modeling the duration of the last IPU syllable yields a 25% relative reduction in classification error over guessing
- Contrary to expectation, the experiments show that the fully automatic setting, Mel-spectral flux (MSF), offers lower error rates than manually verified syllable durations

Impact

- The (next-generation) of dialogue systems should exploit pre-silence lengthening to improve prediction of next-speakership
- We anticipate that the new feature will play a useful role in next-speakership prediction in dialogue as well as in speech rate estimation, improving speech recognition, fluency diagnosis, cognitive load estimation, and others

Supported by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E Prosody in conversation
annah@speech.kth.se, kornel@cs.cmu.edu