

5pSC22. A study of manual articulatory feature-based transcription of conversational speech



Karen Livescu¹, Xuemin Chi¹, Lisa Lavoie¹, Ari Bezman², Nash Borges³, Lisa Yung³

¹MIT, Cambridge, MA USA ²Dartmouth College, Hanover, NH USA ³Johns Hopkins University, Baltimore, MD USA
klivescu@csail.mit.edu



1. Background

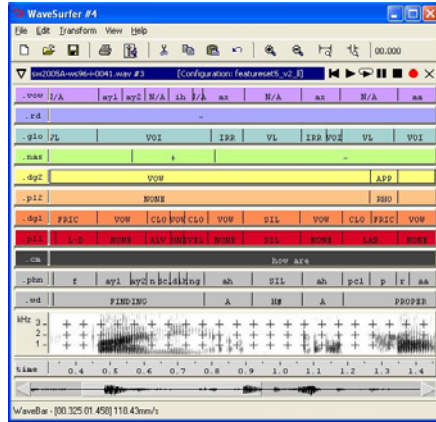
- Motivations: A detailed transcription, including overlapping/reduced gestures**
 - is useful for studying pronunciation variation
 - facilitates the testing of automatic articulatory feature classifiers
- Alternatives:**
 - Narrow phonetic transcription (e.g. Switchboard Transcription Project [1])**
 - may not contain some useful details
 - Physical measurements (e.g. MOCHA [2])**
 - often restricted to scripted tasks
 - have a non-trivial, speaker-dependent relationship to linguistic features
- Goals:**
 - Collect ~100 manually labeled conversational utterances at the level of articulatory features, for use at the 2006 Johns Hopkins Summer Workshop (and beyond)
 - Analyze labeling procedure: speed, inter-transcriber agreement, use of feature tiers
 - Use labeled data to evaluate automatic articulatory feature classifiers

2. Feature set

feature	possible values	description/notes
place 1	LAB, LAB-DEN, DEN, ALV, POST-ALV, VEL, GLO, RHO, LAT, NONE, SIL	Place of forward-most constriction, if any
degree1	VOW, APP, FLAP, FRIC, CLO, SIL	Degree of forward constriction
place 2	LAB-DEN, DEN, ALV, POST-ALV, VEL, GLO, RHO, LAT, NONE, SIL	Place of rear constriction, if any
degree 2	VOW, APP, FLAP, FRIC, CLO, SIL	Degree of rear constriction
nasality	+, -	"+" includes nasalized vowels
glottal state	ST, IRR, VOI, VL, ASP, A+VO	"A+VO" = aspirated and voiced
rounding	+, -	Lip protrusion
vowel	aa, ae, ah, ao, aw1, aw2, ax, axr, ay1, ay2, eh, el, em, en, er, ey1, ey2, ih, ix, iy, ow1, ow2, oy1, oy2, uh, uw, ux, N/A	Vowel quality, if applicable

3. Methods

- Transcribers**
 - Phonetician (author 3) and PhD student in speech research group (author 2)
 - Transcribers participated in designing feature set, transcription interface, methods and conventions
 - Transcriber training minimal, consisting of discussion of practice utterances



Data

- 78 utterances from SVitchboard 1 (small-vocabulary Switchboard 1) [3]
- 9 utterances used in the Switchboard Transcription Project (STP) [1]
- Procedure using WaveSurfer [4]**
 - Basic interface displays waveform, spectrogram, and energy contour
 - Transcribers may play segments, generate waveform blow-ups or use any other WaveSurfer feature
 - 1st pass: Phone-feature hybrid
 - If a segment matches "default" feature values of some phone, the phone label is used; otherwise, feature tiers are used
 - Phone labels are automatically converted to feature values; phone tier is discarded
 - 2nd pass: All-feature transcription
 - Each transcriber views both transcribers' 1st pass annotations side-by-side and makes edits (in her own transcription)
 - 3rd pass: Discussion, clean-up

4. Analysis

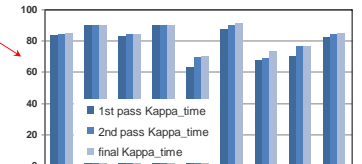
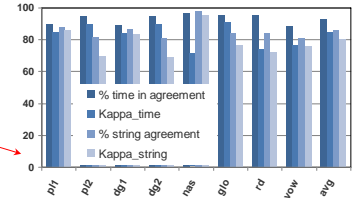
- Transcription time ~1000 x real-time
 - 623 x real-time for 1st pass
 - 335 x real-time for 2nd pass

- Transcriber agreement is high by several measures (note:

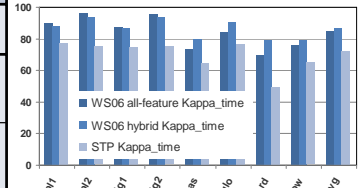
$$Kappa_x = (x_{obs} - x_{chance}) / (x_{total} - x_{chance})$$

- Agreement usually increases slightly in each pass

- Agreement compares favorably with that of STP transcribers (but note: different data sets)
- Preliminary results indicate precision may slightly improve with an all-feature 1st pass



set	transcriber	% canonical
SVitchboard (112.8s)	WS06 transcriber 1	86.2
	WS06 transcriber 2	88.4
STP all-feat (11.4s)	WS06 transcriber 1	81.8
	WS06 transcriber 2	73.5
	STP	84.7
STP hybrid (9.9s)	WS06 transcriber 1	95.3
	WS06 transcriber 2	95.4
	STP	91.6



5. Conclusions

- A new data set** has been collected with human-labeled articulatory feature values
- Transcription approach appears to produce **reliable, detailed annotations**, comparing favorably to a previous effort to produce detailed transcriptions
- In related work, **neural network-based articulatory feature classifiers** have been tested and found to
 - score more poorly against manual transcriptions than against automatic alignments
 - improve after retraining using a feature-based recognizer for alignments
- For additional details/related work, see [4]**

References

- S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus." In Proc. ICSLP 1996.
- A. A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition." In Proc. 5th Seminar on Speech Production: Models and Data, 2000.
- S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small-vocabulary tasks from Switchboard 1." In Proc. Interspeech 2005.
- K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 Johns Hopkins Workshop Final Report." In preparation.