MANUAL TRANSCRIPTION OF CONVERSATIONAL SPEECH AT THE ARTICULATORY FEATURE LEVEL

Karen Livescu[1], *Ari Bezman*[2], *Nash Borges*[3], *Lisa Yung*[3], *Özgür Çetin*[4], *Joe Frankel*[4,5], *Simon King*[5], *Mathew Magimai-Doss*[4], *Xuemin Chi*[1], *Lisa Lavoie*[1]

1. MIT 2. Dartmouth College 3. Johns Hopkins U. 4. ICSI 5. U. Edinburgh

ABSTRACT

We present an approach for the manual labeling of speech at the articulatory feature level, and a new set of labeled conversational speech collected using this approach. A detailed transcription, including overlapping or reduced gestures, is useful for studying the great pronunciation variability in conversational speech. It also facilitates the testing of feature classifiers, such as those used in articulatory approaches to automatic speech recognition. We describe an effort to transcribe a small set of utterances drawn from the Switchboard database using eight articulatory tiers. Two transcribers have labeled these utterances in a multi-pass strategy, allowing for correction of errors. We describe the data collection methods and analyze the data to determine how quickly and reliably this type of transcription can be done. Finally, we demonstrate one use of the new data set by testing a set of multilayer perceptron feature classifiers against both the manual labels and forced alignments.

Index Terms— Speech analysis, speech recognition

1. INTRODUCTION

There has recently been increased interest in the use of articulatory feature (AF) classifiers in automatic speech recognition [1, 2, 3, 4]. One of the potential advantages of this approach is better modeling of coarticulation. A major obstacle, however, is the lack of data labeled with AF values. Classifiers are typically trained and tested on phonetic alignments converted to feature values. However, the actual value of a given feature may differ drastically from its canonical value. Therefore, it may be difficult to model precisely those phenomena one wishes to model with AF-based approaches. This problem is particularly acute in conversational speech, which is characterized by great variability in pronunciation.

One way to address this problem in training AF classifiers is to use an embedded training approach [4]. Using this approach, an initial set of classifiers can be trained based on phone alignments converted to feature values. Then, using the trained classifiers in conjunction with a model of the AF dynamics, a new set of feature alignments (possibly no longer corresponding to canonical phones) is generated. A new set of classifiers is trained using the re-aligned transcriptions, and the process is iterated until some performance threshold is reached. This approach allows for refinement of the original phonetic transcription, but there is still no measure of the quality of the refined transcriptions or of the classifier performance relative to some "ground truth".

There are two other alternatives to automatic phone alignments: physical articulatory measurements and narrow phonetic transcriptions. Physical measurements (e.g. [5] are typically restricted to scripted speech tasks, and, more importantly, can be highly speaker-dependent. Converting such measures to linguistically meaningful features is a challenging task. Manual transcriptions at a narrow phonetic level, such as the Switchboard Transcription Project (STP) [6], contain much of the needed information. However, they can miss some useful information; in the case of STP, examples are unreleased stops and double articulations.

We have begun to address this by collecting a set of manual feature-level transcriptions of conversational utterances. This effort has been conducted in conjunction with the 2006 Johns Hopkins Summer Research Workshop project on articulatory feature-based speech recognition [7]. Our main goal in this paper is to determine whether we can obtain a more detailed transcription at the feature level than would be afforded by, for example, converting STP transcriptions to features, while maintaining reasonable transcriber agreement. In the following sections, we describe the development of a feature set and transcription interface intended to enable detailed and reliable labeling; the generation of manual labels for 78 utterances drawn from SVitchboard [8], a small-vocabulary subset of Switchboard, and an additional 9 utterances drawn from the STP set; analysis of the data for inter-transcriber agreement and prevalence of canonical vs. non-canonical labels; and an example use of the data for testing a set of AF classifiers.

2. FEATURE SET

At the outset of this project, we started with a set of features similar to the ones used in the International Phonetic Alphabet to distinguish phones. In the process of developing the transcription procedure, we refined the feature set to better describe effects such as double articulations. The resulting feature set is shown in Table 1. It allows for up to two simultaneous constrictions, each of which is described by place and degree (or manner) features: **pl1** and **dg1** are the place and degree of the forward-most constriction, i.e. the one closest to the lips, if any; **pl2** and **dg2** refer to a second constriction, if any, farther back in the vocal tract. We note that **dg1** and **dg2** are not exactly physical degrees of constriction; for example, the same constriction can result in a fricative or approximant, depending on the pressure behind the constriction. In addition, we have collapsed the traditional height and front-back vowel features into a single vowel quality tier; we found that this was quicker to label than the two separate vowel features, without any apparent information loss.

This material is based upon work supported by the National Science Foundation under Grant No. 0121285. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was partly supported by the Swiss National Science Foundation through the research network IM2. We are grateful to the entire JHU WS06 articulatory feature-based speech recognition team for comments and suggestions, and to the JHU Center for Language and Speech Processing for facilitating this research.

Feature	Values
pl1	labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal, rhotic, lateral, none, silence
dg1	vowel (no constriction), approximant, flap, fricative, closure, silence
pl2	As in **pl1**, minus labial
dg2	As in **dg1**
nas (nasality)	+, -
glo (glottal state; includes some supra-glottal information)	voiced, voiceless, glottal stop, irregular pitch periods, aspiration, aspiration + voicing
rd (lip rounding)	+, -
vow (vowel quality)	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 (not a vowel)

Table 1. *Feature set used in transcriptions. For diphthongs, [label]1 refers to the starting state and [label]2 refers to the ending state.*

3. METHODS

The main data set consists of 78 utterances drawn from SVitchboard [8], a small-vocabulary subset of Switchboard. Of these, 33 were hand-selected (by author 1) and 45 were randomly drawn from all 5- to 14-word SVitchboard utterances. An additional 9 utterances drawn from the STP data set were also transcribed, for comparisons with STP. The data was labeled by two transcribers, a phonetician and a graduate student with experience in speech research (authors 10 and 9). The transcribers were not formally trained, but rather labeled and discussed practice utterances while participating in refinement of the feature set and transcription interface.

The transcription interface was based on KTH's WaveSurfer [9], a highly configurable sound analysis and annotation tool. Figure 1 shows a screen shot. Besides the feature tiers, the transcribers were provided wide-band and narrow-band spectrograms and a plot of the signal power. There was no time restriction and the transcribers could use all sources of information available in WaveSurfer (e.g. listening to arbitrary waveform segments, generating spectral slices or waveform blow-ups, and modifying the spectrogram parameters). The procedure for transcription was as follows:

- *Initialization*. Transcribers were provided initial word and phone alignments, to be modified during transcription. The word alignments were those published by Mississippi State University [10]. The phone alignment was a manual alignment, within the given word boundaries, to a *dictionary* pronunciation of each word,[1] done by author 1. The initial and final silence feature values were filled in automatically.
- *First pass*. Each transcriber labeled the utterances using a hybrid phone-feature labeling: For any segment that could be described as a canonical phone, the phone label was used; otherwise, the feature tiers were used. These transcriptions were then automatically converted to all-feature transcriptions before proceeding. From this point on the phone tier was dropped and only the feature tiers were used.
- *Second pass*. Each transcriber compared her transcriptions to the other transcriber's and corrected any errors in her own labeling (but not disagreements). For this purpose, a different WaveSurfer configuration was used, with the two transcriptions viewed together.
- *Third pass*. Transcribers discussed disagreements and, possibly, modified their transcriptions.

[1] Based on a dictionary from the MIT Spoken Language Systems Group.

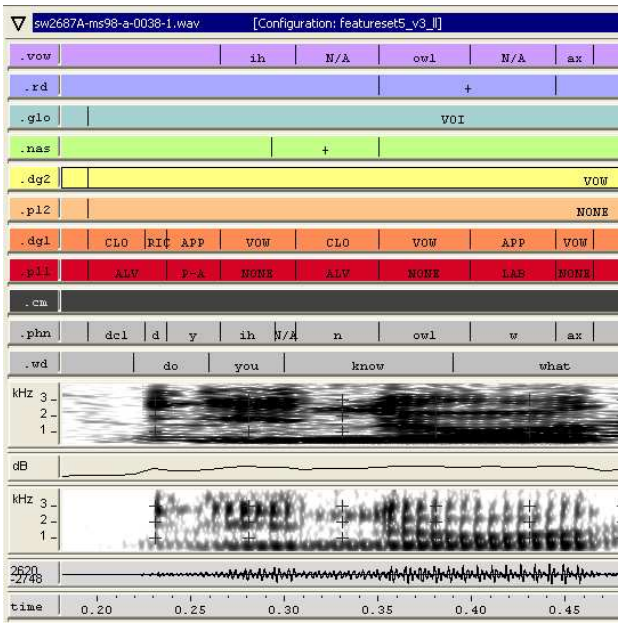

Fig. 1. *Screen shot of the transcription interface.*

The phone-feature hybrid was chosen for the first pass because transcribers found it tedious to enter feature values for perceptually canonical phones. The transcribers were provided a set of labeling conventions and guidelines.[2] Some aspects of articulation are, of course, not discernible from the signal (such as a tongue constriction during a [p] closure); however, only those aspects that are acoustically salient need be labeled correctly to satisfy the above guideline. Another way to view this is that the transcriptions should reflect what we would like automatic AF classifiers to be able to detect. Averaged over the 78 SVitchboard utterances excluding initial and final silences (a total of 119s of speech), the speed of transcription was 623 times real-time for the first pass and 335 times real-time for the second pass.

The nine STP utterances were treated somewhat differently. In order to analyze the effect of using the hybrid phone-feature format in the first pass, four of the STP utterances (9.9s of non-silence speech) were transcribed as described above, and the remaining five (11.4s) were transcribed using an all-feature format in the first pass. We found that, for these utterances, the first-pass real-time factor was 328 for the phone-feature hybrid transcriptions and 799 for the all-feature transcriptions.

4. ANALYSIS

To measure the reliability and usefulness of our approach, we examine both inter-transcriber agreement and the degree to which the transcribers used the available feature tiers. Figure 2 shows several measures of transcriber agreement computed on the SVitchboard utterances. The time-weighted agreement is the proportion of the time

[2] Available at `http://people.csail.mit.edu/klivescu/twiki/bin/view.cgi/WS06/TranscriptionNotes`

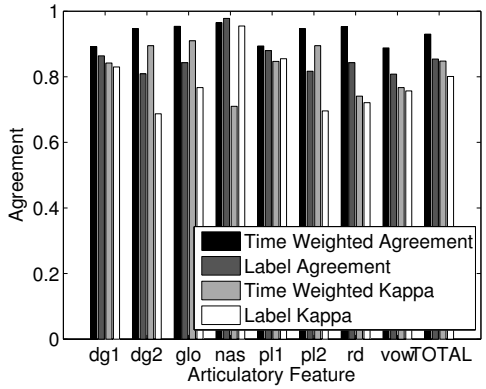

Fig. 2. *Transcriber agreement measures on SVitchboard utterances.*

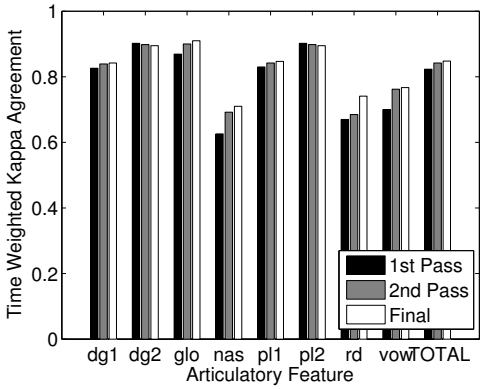

Fig. 3. *Time-weighted Kappa for each pass through the manually transcribed SVitchboard utterances.*

the transcribers agreed. In order to account for chance agreement, the corresponding time-weighted Kappa statistic was computed:

$$\text{Kappa}_{time} = \frac{t_{\text{obs}} - t_{\text{chance}}}{t_{\text{total}} - t_{\text{chance}}}, \quad (1)$$

where t_{obs} is the observed time in agreement, t_{chance} is the time in agreement expected by chance, and t_{total} is the total time. We also measure the label agreement, or string accuracy, and the associated Kappa statistic using alignments of the transcribers' label strings. Figure 5 shows the inter-transcriber confusion matrix for **dg1**, in terms of time-weighted agreement. For the remaining figures, we consider only the time-weighted Kappa statistic.

To determine how much of the agreement we observe is due to the multi-pass transcription refinement, we compare agreement across multiple passes (Figure 3). While the agreement usually improves from the first to the final pass, it is not a dramatic difference.

Figure 4 compares our inter-transcriber agreement on STP utterances to that of STP's transcribers. For this purpose, we converted the STP labels to our feature set. Note that the STP agreement data comprises a different set of utterances from the 9 we have transcribed; therefore, these numbers should be viewed as only a rough guideline of the overall agreement differences.

The main conclusions are that our transcriptions have a high level of inter-transcriber agreement, even in the first pass, and compare favorably with the agreement statistics of STP.

Next we measure the extent to which the transcribers used the extra flexibility of the feature tiers that is not available in a phonetic transcription. Table 2 shows the percentage of the time, excluding silences, for which the transcribers used canonical feature configurations. We note that by "canonical", we mean any feature configuration corresponding to a phone label, not necessarily the phone label in the dictionary pronunciation. The results indicate that the WS06 transcribers were more likely than STP's transcribers to use non-canonical feature configurations when using an all-feature format in the first pass, and less likely to do so when using a phone-feature hybrid format in the first pass. Although this measure is based on a very small data set, it suggests that future transcription efforts may benefit from investigating improvements to the transcription interface to encourage greater use of the feature tiers.

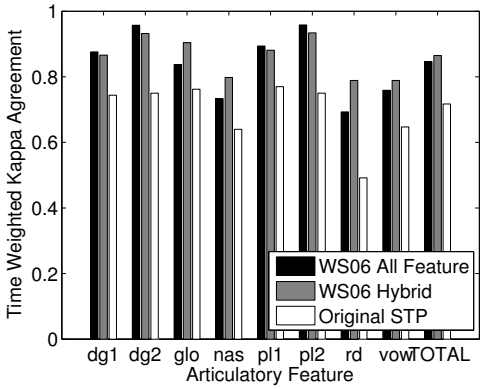

Fig. 4. *Time-weighted Kappa for transcriptions of STP utterances by WS06 and STP transcribers. The WS06 transcriptions have been split into those 5 utterances that had all-feature first-pass transcriptions and the 4 with hybrid (phone-feature) first-pass transcriptions.*

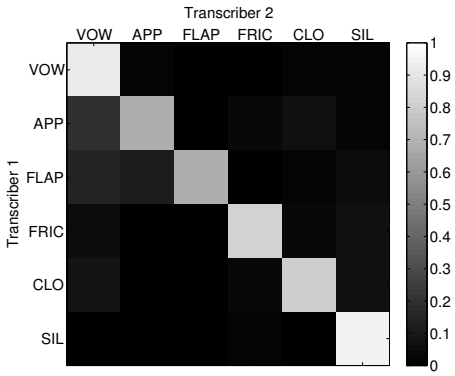

Fig. 5. *Inter-transcriber confusion matrix for* **dg1** *in the 78 manually transcribed SVitchboard utterances.*

5. CLASSIFIER PERFORMANCE

As an example use of the newly collected data, we test a set of articulatory feature classifiers against both the manual labels and forced phone alignments converted to feature labels. The classifiers are multilayer perceptrons (MLPs), one per feature, trained on a set of automatically generated phone alignments converted to feature la-

Set	transcriber	% canonical
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
SVitchboard (112.8s)	WS06 transcriber 1	86.2
	WS06 transcriber 2	88.4

Table 2. *Percent of the time for which canonical feature configurations are used, for different utterance sets and transcribers. In the leftmost column, numbers in parentheses represent the total (non-silence) duration of the utterance set.*

bels.[3] Table 3 shows the MLPs' performance on the 78 manually-transcribed SVitchboard utterances. The accuracies relative to the WS06 transcribers are very similar, and are lower than the accuracies relative to the forced alignments, as is perhaps expected since the MLPs were trained on forced alignments. As an example of the types of confusions observed, Figure 6 displays the confusion matrices for **dg1**, relative to both the forced alignments and Transcriber 1's labels. For additional information, see [7, 11].

Reference	Accuracy					
	dg1	pl1	glo	nas	rd	vow
Forced align.	78/41	78/41	87/53	97/95	94/91	83/74
Transcriber 1	74/47	74/47	86/57	95/94	92/90	78/73
Transcriber 2	73/47	72/47	86/57	94/93	92/90	77/71

Table 3. *Accuracies (in %) of MLP AF classifiers evaluated against forced alignments and against the WS06 transcribers. The format of the table cells is "[MLP accuracy]/[chance accuracy]", where chance accuracy is the accuracy that would be obtained by choosing the most common label value in the reference data. There are no results for* **dg2** *and* **pl2***, since these are always "none" or "silence" in the automatically generated training labels.*

6. CONCLUSIONS

We have described a new set of feature-level manual transcriptions of conversational speech and an approach for generating such transcriptions. The data will be publicly available for download from `http://people.csail.mit.edu/klivescu/WS06AFSR`.

We have analyzed our transcription approach in terms of transcriber agreement and the preponderance of canonical vs. noncanonical labels. We have found that transcriptions of SVitchboard utterances can be done in roughly 1000 times real-time and can have high inter-transcriber agreement. Our analysis also suggests that the phone-feature format for first-pass transcriptions may have reduced the use of non-canonical labels. This is a point that may require further study with a larger set of utterances, and may indicate that future similar transcription efforts can benefit from an interface that facilitates more convenient use of the feature tiers. Finally, we have found that, as may be expected, AF classifiers trained on phonetic forced alignments have lower accuracy when tested against manual transcriptions. This suggests that testing classifiers relative to forced alignments may overestimate their accuracies. We anticipate that the generated transcriptions will be a useful aid in refining AF classifiers and embedded training methods.

[3] 1776 hours of training and 225 hours of cross-validation data from the Fisher, Switchboard Cellular, Switchboard Credit-card, and Switchboard 2 corpora. We are grateful to SRI for providing the alignments.

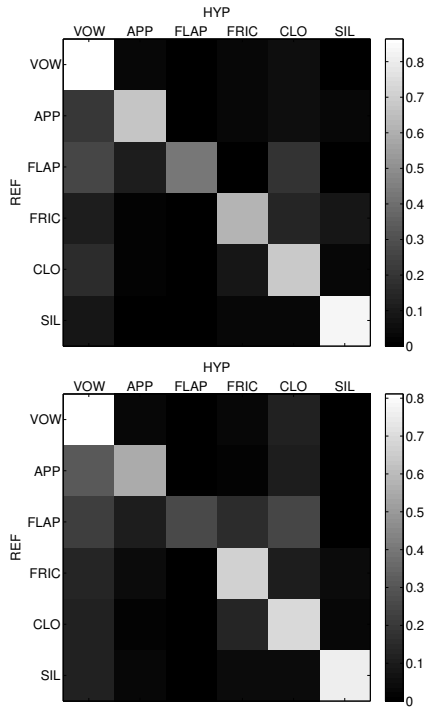

Fig. 6. *Confusion matrices of MLP classifier outputs (HYP) for* **dg1***, relative to forced phone alignments converted to feature values (REF, top) and to Transcriber 1 (REF, bottom).*

7. REFERENCES

[1] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2000.

[2] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP*, 2002.

[3] M. Hasegawa-Johnson et al., "Landmark-based speech recognition," in *ICASSP*, 2005.

[4] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *Proc. IEICI Beyond HMM Workshop*, 2004.

[5] 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.

[6] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. ICSLP*, 1996.

[7] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," in *ICASSP*, 2007.

[8] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from Switchboard 1," in *Proc. Interspeech*, 2005.

[9] WaveSurfer, "http://www.speech.kth.se/wavesurfer/," .

[10] N. Ganapathiraju, A. Deshmukh, A. Gleeson, A. Hamakera, and J. Picone, "Resegmentation of SWITCHBOARD," in *Proc. ICSLP*, 1998.

[11] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: JHU Summer Workshop Final Report," Technical report, Johns Hopkins University Center for Language and Speech Processing, 2007, in preparation.