AUTOMATIC PHONETIC TRANSCRIPTION OF SPONTANEOUS SPEECH (AMERICAN ENGLISH)

Shuangyu Chang, Lokendra Shastri and Steven Greenberg

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
{shawnc, shastri, steveng}@icsi.berkeley.edu

ABSTRACT

An automatic transcription system has been developed to label and segment phonetic constituents of spontaneous American English without benefit of a word-level transcript. Instead, special-purpose neural networks classify each 10-ms frame of speech in terms of articulatory-acoustic-based phonetic features and the feature clusters are subsequently mapped to phonetic-segment labels using multilayer perceptron networks. The phonetic labels generated by this system are 80% concordant with the labels produced by human transcribers and the segmental boundaries deviate from manual segmentation by an average of 11 ms. The automatic transcription system thus generates phonetic labels and segmentation comparable in quality to those produced by human transcribers, and therefore may prove useful for phonetic annotation of novel linguistic corpora, as well as facilitating development of pronunciation models for automatic speech recognition systems.

1. INTRODUCTION

Current-generation speech recognition (ASR) systems generally rely on automatic-alignment procedures to train and develop phonetic-segment models. Although these automatically generated alignments are designed to approximate the actual phones contained in an utterance they are often erroneous in terms of both phonetic identity and segmentation boundaries. Over forty percent of the phonetic labels generated by state-of-the-art automatic alignment systems differ from those generated by phonetically trained individuals [3]. Moreover, the boundaries generated by these automatic alignment systems differ by an average of 32 ms (40% of the mean phone duration) from the hand-labeled material [3]. The quality of automatic labeling and segmentation is potentially of great significance for large-vocabulary ASR system performance since word-error rate appears to be largely dependent on the accuracy of phone recognition and segmentation [3]. Moreover, a substantial reduction in word-error rate is, in principle, achievable when phone recognition is both extremely accurate and tuned to the phonetic composition of the recognition lexicon [5]. An accurate method of automatic phonetic transcription could potentially facilitate development of ASR systems for novel material, both within and across languages, as well as increase robustness with respect to acoustic interference and variation in speaking style and pronunciation.

The current study describes an automatic system for automatic labeling of phonetic segments (ALPS) in utterances drawn from a corpus of spontaneous American English (OGI Numbers95). The performance of the ALPS system is comparable in accuracy and reliability to that of human transcribers and is achieved without using a word-level transcript (as automatic-alignment systems require). The system's initial classification (using special-purpose neural networks) is based on recognizing articulatory-acoustic phonetic features (AFs) rather than phones. These phonetic features are subsequently mapped to phonetic-segment labels using a separate set of neural networks that also form the basis of delineating the segmental boundaries.

2. TRANSCRIPTION SYSTEM OVERVIEW

The speech signal is processed in several stages (cf. Figure 1). First, a power spectrum is computed every 10 ms (over a 25-ms window) and this spectrum partitioned into quarter-octave channels between 0.3 and 4 kHz. The power spectrum is logarithmically compressed in order to preserve the general shape of the spectrum distributed across frequency and time (an example of which is illustrated in Figure 2 for the manner-of-articulation feature, *vocalic*).

An array of independent, temporal flow neural networks (cf. Section 4) classify each 25-ms frame along five articulatory-based, phonetic-feature dimensions: (1) place and (2) manner of articulation, (3) voicing, (4) lip-rounding and (5) front-back articulation (for vocalic segments). A separate class was derived for “silence” (labeled as “null” in each feature dimension). These phonetic-feature labels are combined and serve as the input to a multilayer perceptron (MLP) network that performs a preliminary classification of phonetic identity (e.g., [f] [ay] [v]). The output of these networks is processed by a Viterbi-like decoder to produce a sequence of phonetic-segment labels along with boundary demarcations associated with each segment.

3. CORPUS MATERIALS

The ALPS transcription system was evaluated using spontaneous speech material from the Numbers95 corpus [1], collected and phonetically annotated (i.e., labeled and segmented) at the Oregon Graduate Institute. This corpus contains the numerical portion (mostly street addresses and phone numbers) of thousands of telephone dialogues and possesses a lexicon of 37 words and an inventory of 29 phonetic segments. The speakers contained in the corpus are of both genders and represent a wide range of dialect regions and age groups.

The ALPS system was trained on ca. 2.5 hours of material and a separate 15-minute, cross-validation set was used for training the networks and setting the appropriate threshold parameters. Testing and evaluation of the transcription systems was performed on an independent set of ca. one hour's duration.

4. TEMPORAL FLOW MODEL NETWORKS

In the ALPS system initial classification of articulatory-acoustic features is performed by temporal flow model (TFM) networks [10]. These networks support arbitrary link connectivity across multiple layers of nodes, admit feed-forward, as well as recurrent links and allow variable propagation delays to be

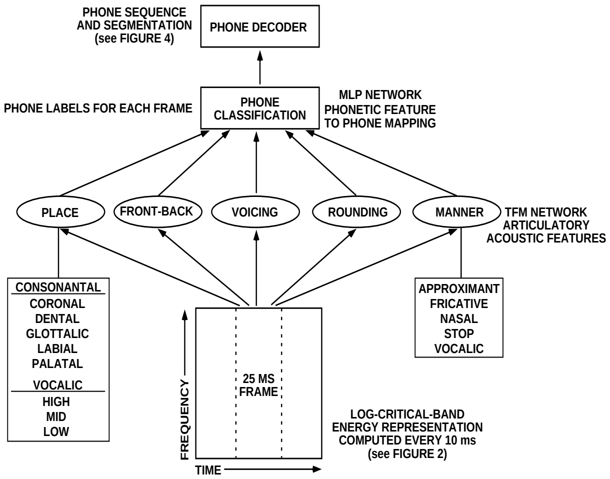

Figure 1. Schematic description of the ALPS automatic transcription system using articulatory-acoustic features to label and segment phones in spontaneous speech.

associated with each of the links. The recurrent links in TFM networks provide an effective means of smoothing and differentiating signals, as well as detecting the onset (and measuring the duration) of specific features. Using multiple links with variable delays allows a network to maintain an explicit context over a specified window of time and thereby makes it capable of performing spatio-temporal feature detection and pattern matching. Recurrent links, used in tandem with variable propagation delays, provide a powerful mechanism for simulating certain properties (such as short-term memory, integration and context sensitivity) essential for processing time-varying signals such as speech. TFM-based networks have been shown to perform as well as, if not better than, standard neural networks (such as MLPs) using an architecture that is far more efficiently constructed (cf. Table 1). In the past, TFM networks have been successfully applied to a wide variety of pattern-classification tasks including phoneme classification [9], optical character recognition [8] and syllable segmentation [7]. The TFM networks used to classify articulatory features in the ALPS system possess between 3,000 and 8,000 adjustable weights.

5. SPECTRO-TEMPORAL PROFILES

The architecture of the TFM networks used for classification of the articulatory acoustic features was developed using a three-dimensional representation of the log-power-spectrum distributed across frequency and time that incorporates both the mean and variance of the energy distribution associated with multiple (typically, hundreds or thousands of) instances of a specific phonetic feature or segment derived from the phonetically annotated, OGI Stories-TS corpus [1]. Each phonetic-segment class was mapped to an array of articulatory phonetic features, and this map used to construct the spectro-temporal profile (STeP) for a given feature class. For example, the STeP for the manner feature, *vocalic* (cf. Figure 2), was derived from a summation of all instances of vowel segments in the corpus. The STeP extends 500 ms into the past, as well as 500 ms into the future relative to the reference frame, t_0, thereby spanning an interval of 1 second. This extended window of time is designed to accommodate co-articulatory context effects. The frequency dimension is partitioned into quarter-octave channels. The variance associated with each component of the STeP is color-coded and identifies those regions which most clearly exemplify the energy-modulation patterns across time and frequency associated with the feature of interest (cf. Figure 2) and can be used to adjust the network connectivity in appropriate fashion.

Network Type	TFM + MLP	MLP	MLP
Context (Frames)	9	9	19
Hidden Units in MLP	400	800	800
Total Parameters	130,600	128,800	241,000
Frame Accuracy (%)	79.4	73.4	79.4

Table 1. Frame-level phonetic-segment classification accuracy for different neural-network architectures and context lengths.

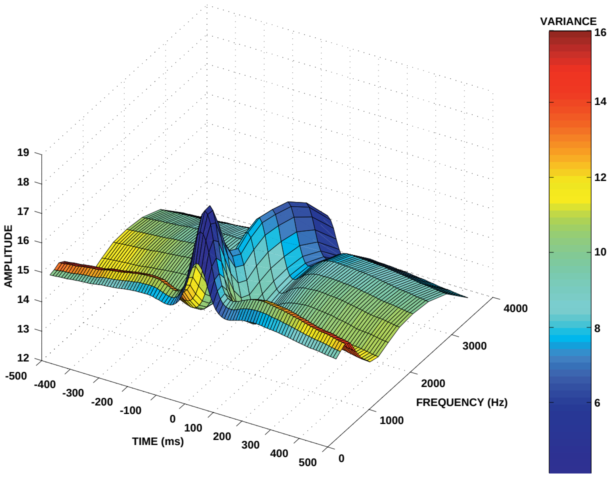

Figure 2. A spectro-temporal profile of the phonetic feature, *vocalic*, derived from the superposition of thousands of instances of this feature in the OGI Stories-TS corpus [1].

6. PHONETIC-SEGMENT DECODING

An MLP network, possessing a single hidden layer of 400 units, was used to map the phonetic features derived from the TFM networks onto phonetic-segment labels. The input to the MLP used a context window of 9 frames (105 ms). The output of this MLP contains a vector of phone-probability estimates for each 10-ms frame. This matrix of phonetic-segment probabilities is converted into a linear sequence of phone labels and segmentation boundaries via a decoder. A hidden-Markov-model (HMM) was applied to impose a minimum-length constraint on the duration associated with each phonetic-segment (based on segmental statistics of the training data), and a Viterbi-like decoder used to compute the sequence of phonetic segments over the entire length of the utterance. This bipartite decoding process is analogous to that used for decoding word sequences in automatic speech recognition systems. However, in the present application, the “lexical” units are phones, rather than words, and the “words” contain clusters of articulatory features rather than phones. It is also possible to convert the frame-level, phonetic-feature data into phone sequences by using a threshold model derived from the statistical characteristics of a separate (validation) data set. In this instance a minimum-duration constraint is imposed as a means of smoothing the output of the phone-selection process. Both the threshold and HMM-based approaches produce equivalent results (Table 1).

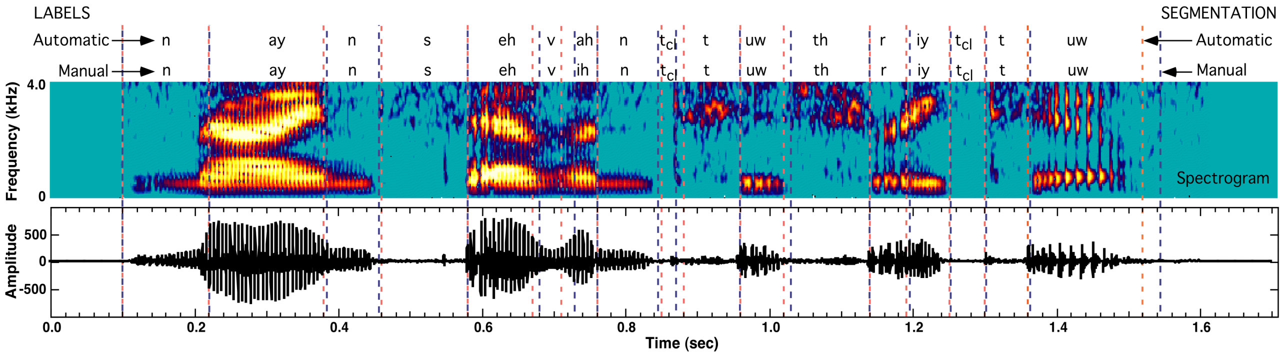

Figure 3. The labels and segmentation generated by the ALPS transcription system for the utterance "Nine, seven, two, three, two" are compared to those produced manually. The top row shows the phone sequence produced by the ALPS system. The tier directly below is the phone sequence produced by a human transcriber. The spectrographic representation and waveform of the speech signal are shown below the phone sequences as a means of evaluating the quality of the phonetic segmentation. The manual segmentation is marked in purple while the automatic segmentation is illustrated in orange.

A separate TFM neural network was used to compute the precise location of the segment boundaries based on the matrix of phone probabilities distributed across frames. The identity of the phone segments combined with their associated boundaries form the output of the system's phonetic transcription (cf. Figure 3).

7. EVALUATION OF THE ALPS SYSTEM

7.1 Articulatory-Acoustic Feature Classification

The accuracy of articulatory-acoustic feature classification ranges between 79% (place of articulation) and 91% (voicing) (Table 2), and is comparable or superior to the performance obtained by Kirchhoff [4] using MLP networks. In the current study AF classification was performed by manually tuned TFM networks, based on information contained in the STePs associated with each of the relevant articulatory-based features.

7.2 Phonetic-Segment Classification

Table 1 illustrates the capability of the ALPS system to map articulatory-acoustic features onto phonetic-segment labels using an MLP network operating on the AF output of the TFM networks. The table compares the performance of this hybrid system with that of two different MLP-based phone classification systems. The TFM/MLP system significantly outperforms the standard MLP phone classifier (which uses 9 frames of context) and is comparable in classification accuracy to an MLP using 19 frames (205 ms) of context. However, the TFM/MLP system achieves this level of performance with less than half of the parameters required by the MLP classifier alone.

The frame accuracy of phonetic classification associated with the hybrid TFM/MLP system can be increased from 79.4% to 82.5% by *reducing* the temporal resolution of the inputs to the TFM and MLP neural networks by a factor of four (but not by factors of two or eight) and then combining the output with that of an MLP processing the original (10-ms) resolution features. This experiment suggests that there is information contained in ca. 40-ms-length segments of particular importance for phonetic classification.

Phonetic-Feature Parameter				
Place	**Front/Back**	**Voicing**	**Rounding**	**Manner**
78.8	**83.4**	**91.1**	**85.6**	**84.4**

Table 2. Frame-level accuracy (percent correct) of phonetic feature classification for the ALPS transcription system.

7.3 Temporal Location of Phonetic Labeling Errors

It is of interest to ascertain the frame location of phonetic-segment classification errors as a means of gaining insight into the origins of mislabeling this material. Specifically, it is important to know whether the classification errors are randomly distributed across frames or are concentrated close to the segment boundaries. The data illustrated in Figure 4 indicate that a disproportionate number of errors are concentrated near the phonetic-segment boundaries in regions inherently difficult to classify accurately as a consequence of the transitional nature of phonetic information in such locations. Nearly a third of the phone classification errors are associated with boundary frames accounting for just 17% of the utterance duration. The accuracy of phone classification is only 61% in the boundary frames, but rises to 80% or higher for frames located in the central region of the phonetic segment.

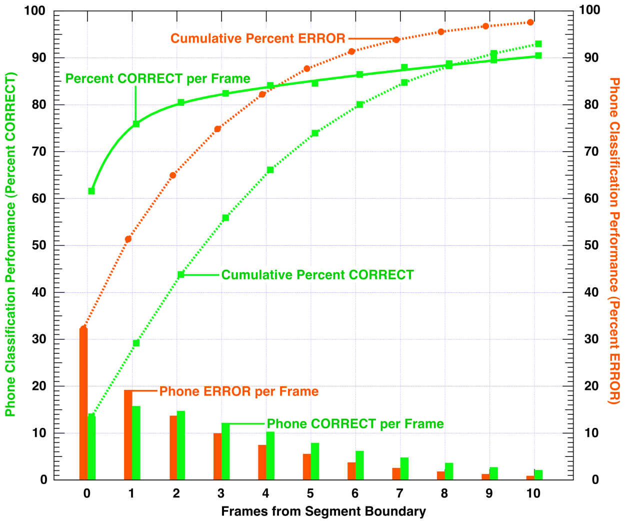

Figure 4. Phonetic-segment classification performance as a function of frame (10 ms) distance from the manually defined phonetic-segment boundary. Contribution of each frame to the total number of correct (green) and incorrect (orange) phonetic segments classified by the ALPS system is indicated by the bars. The cumulative performance over frames is indicated (dashed lines), as is the percent correct phone classification for each frame (green squares, with a double-exponential, solid-line fit to the data).

Procedure	Substitutions	Deletions	Insertions	Total
HMM	8.1	6.4	4.9	19.3
Threshold	6.9	8.4	4.3	19.5

Table 3. Percent error associated with phonetic-label decoding, partitioned by error type, for two different decoding methods.

7.4 Phonetic-Segment Decoding

The performance of two separate methods of decoding phonetic sequences (one based on HHMs, the other on a threshold model - cf. Section 6) are compared in Table 3. The decoding techniques produce essentially equivalent results. However, the threshold model produces slightly fewer substitution errors than the HMM procedure and may therefore be of greater utility under certain conditions where fidelity of transcription is of prime concern.

7.5 Phonetic Segmentation

The accuracy of phonetic segmentation can be evaluated by computing the proportion of times that a phonetic segment onset is correctly identified ("hits") by the ALPS system relative to the instances where the phone onset (as marked by a human transcriber) is located at a different frame ("false alarms"). The data in Table 4 indicate that the ALPS system matches the segmentation of human transcribers precisely in ca. 40% of the instances. However, automatic segmentation comes much closer to approximating human performance when a tolerance level of more than a single frame is allowed (76-84% concordance with manual segmentation). The average deviation between the manual and automatic segmentation is 11 ms, an interval that is ca. 10% of the average phone duration in the Numbers95 corpus.

8. DISCUSSION AND CONCLUSIONS

The ALPS transcription system possesses certain advantages over other methods of automatically labeling phonetic segments in spontaneous speech. It does not require a word-level transcript (as is the case with forced-alignment procedures and other techniques, such as MAUS developed at the University of Munich [6]). In addition, the ALPS system is likely to be relatively robust in the presence of acoustic interference [4] and speaking-style variation [2] since the initial classification is based on a relatively small number of articulatory acoustic features rather than on phones. Articulatory-acoustic features also provide a means of more accurately delineating the phonetic composition of spontaneous material since speech is rarely spoken in perfectly canonical fashion. Often, specific articulatory-acoustic features are either absent or their time-course deviates from that of associated features within a phonetic segment. Because most of the articulatory features used to develop the ALPS system are also present in most other languages of the world, it is inherently cross-linguistic in capability and extensible to other corpora.

To date the ALPS system has been applied only to a single corpus of relatively restricted phonetic composition. In the future we intend to apply the system to more complex corpora of American English, as well as to corpora of other languages.

Frame Tolerance	Hits	False Alarms
±1 (10 ms)	38.4	58.5
±2 (20 ms)	76.0	20.9
±3 (30 ms)	83.7	13.2

Table 4. Accuracy of phonetic segmentation as a function of the temporal tolerance window and partitioned into error type (hits/false alarms).

ACKNOWLEDGEMENTS

The research described in this study was supported by the U.S. Department of Defense and the National Science Foundation (Learning and Intelligent Systems Initiative). The authors wish to thank Climent Nadeu for his helpful comments on an earlier version of this paper.

REFERENCES

[1] Cole, R., Fanty, M., Noel, M. and Lander, T. "Telephone speech corpus development at CSLU," *Proc. Int. Conf. Spoken Lang. Proc.*, 1994.

[2] Deng, L., Ramsay, G. and Sun., D. "Production models as a structural basis for automatic speech recognition," *Speech Communication*, 22: 93-112, 1997.

[3] Greenberg, S., Chang, S., and Hollenback, J. "An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems," *Proc. NIST Speech Transcription Workshop*, 2000.

[4] Kirchhoff, K. *Robust Speech Recognition Using Articulatory Information*, Ph.D. Thesis, University of Bielefeld, 1999.

[5] McAllaster, D., Gillick, L., Scattone, F. and Newman, M. "Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch," *Proc. Int. Conf. Spoken Lang. Proc.*, 1998.

[6] Schiel, F. "Automatic phonetic transcription of non-prompted speech," *Proc. Int. Cong. Phon. Sci.*, pp. 607-610, 1999.

[7] Shastri, L. Chang, S. and Greenberg, S. "Syllable detection and segmentation using temporal flow model neural networks. *Proc. Int. Cong. Phon. Sci.*, pp. 1721-1724, 1999.

[8] Shastri, L., and Fontaine, L. "Recognizing hand-written digit strings using modular spatio-temporal neural networks," *Connection Science* 7 (nos. 3 and 4), 1995.

[9] Watrous, R. "Phoneme discrimination using connectionist networks," *J. Acoust. Soc. Am.*, 87: 1753-1772, 1990.

[10] Watrous, R. and Shastri, L. "Learning phonetic features using connectionist networks," *Proc. Tenth Int. Joint Conf. Artificial Intelligence*, pp. 851-854, 1987.