ROBUST AUTOMATIC TIME ALIGNMENT OF ORTHOGRAPHIC TRANSCRIPTIONS WITH UNCONSTRAINED SPEECH

Barbara Wheatley, George Doddington,† Charles Hemphill, John Godfrey,
Edward Holliman, Jane McDaniel, and Drew Fisher

Texas Instruments, P.O. Box 655474, MS 238, Dallas, Texas 75265

ABSTRACT

This paper presents a method for automatic time alignment of orthographically transcribed speech using supervised speaker-independent automatic speech recognition based on the orthographic transcription, an on-line dictionary, and HMM phone models. This method successfully aligns transcriptions with speech in unconstrained 5 to 10 minute conversations collected over long-distance telephone lines. It requires minimal manual processing and generally produces correct alignments despite the challenging nature of the data. The robustness and efficiency of the method make it a practical tool for very large speech corpora.

1. PROBLEM

Time alignment of orthographic transcriptions with speech data is essential for the effective use of speech corpora such as the SWITCHBOARD corpus now being created at Texas Instruments [1]. This corpus will consist of 2500 conversations conducted over long-distance telephone lines. Each conversation contains from 5 to 10 minutes of spontaneous speech between two participants, with each channel recorded separately so that the speakers can be separated or combined at will. To make this corpus suitable for research in tasks such as speaker identification and large-vocabulary speech recognition, we are determining the timing of each speaker's turn and the timing of each word to sufficient precision to localize words of interest to a small segment of the signal.

The time alignment task for this corpus involves a number of challenging aspects: the large size of each file (the corpus average is 7.7 minutes); conversational phenomena such as simultaneous speech; the wide range of linguistic material occurring in unconstrained conversations; the bandwidth and noise characteristics of long-distance telephone speech; cross-channel echo; and variations in channel synchronization.

Previous work on automatic time alignment of speech has typically made use of independently-supplied phonetic transcriptions [2]. Other work has determined time alignment directly from orthographic transcriptions [3], but has differed in scope and goals, attempting to achieve correct alignment at the phonetic level on more tractable data (the TIMIT database [4]). Our data are significantly more challenging, and our goals more modest.

2. METHOD

The time alignment procedure involves four steps: a) creating a supervision grammar from the orthographic transcription; b) generating a grammar for each word in the transcription, based on an on-line dictionary and phonological rule set; c) executing supervised recognition, and d) extracting the timing information from the recognition output.

Each conversation in the Switchboard corpus is transcribed orthographically with speaker turns indicated. Instances where both participants speak simultaneously are also identified in the transcription. From the orthographic transcription, we automatically generate a finite-state grammar uniquely characterizing the observed word sequence. This grammar dictates a strict linear progression through the text except for simultaneous speech, as discussed below.

Nonspeech sounds, such as breath noises and laughter, are also indicated in the transcription but are omitted from the conversation-level grammar. However, silence and nonspeech sounds are accommodated in the grammar in the form of self-loops at each node, i.e., initially, finally, and between each pair of words. Explicit recognition models, trained on the Voice Across America long-distance telephone speech database [5], are used for silence, inhalation, exhalation, and lip smack. All other nonspeech sounds are not explicitly modeled, but are instead accommodated through the use of a score threshold, which automatically classifies as nonspeech any input frame not sufficiently close to any candidate recognition model.

†Now at SRI International, Menlo Park, California.

0-7803-0532-9/92 $3.00 © 1992 IEEE

```
start(BYE_f).                start(CITIES_f).
BYE_f ---> b_f, z_0_f.       CITIES_f ---> s_f, z_0_f.
z_0_f ---> ay_f, z_2_f.      z_0_f ---> ih_f, z_2_f.
z_2_f ---> "".               z_2_f ---> dx_f, z_4_f.
start(CAN_f).                z_2_f ---> t_f, z_4_f.
CAN_f ---> k_f, z_2_f.       z_4_f ---> iy_f, z_3_f.
z_2_f ---> ae_f, z_1_f.      z_3_f ---> z_f, z_5_f.
z_2_f ---> en_f, z_3_f.      z_5_f ---> "".
z_2_f ---> ix_f, z_1_f.
z_1_f ---> n_f, z_3_f.
z_3_f ---> "".
```

Figure 1. Examples of word models for a female speaker. These models allow one pronunciation for BYE, three for CAN, and two for CITIES, including the medial flap pronunciation derived by rule.

For each word occurring in the conversation, we generate a finite-state grammar representing one or more pronunciations. The pronunciations are obtained from a 240,000-entry on-line dictionary. A separate path through the word-level grammar is generated for each alternate pronunciation represented in the dictionary. In addition, alternate paths are added for optional variants derived by applying phonological rules such as alveolar stop flapping. Examples are shown in Figure 1.

All of the steps in conversation-level and word-level grammar creation are fully automated. The sole manual operation in the time-alignment procedure is adding new words to the dictionary as they occur in conversations. Initially, each conversation required the addition of 20-25 words, but this rate has decayed rapidly. Currently, the dictionary has been augmented for 1153 conversations at an overall rate of 2.6 words per conversation. Most of the added words are proper nouns.

Word pronunciations are realized in terms of a set of context-independent phone models. These phone models are continuous-density HMMs that have been trained for speaker-independent recognition of long-distance telephone speech on 1000 phonetically-balanced sentences (based on TIMIT sentences) in the Voice Across America database. Each phone has two variants, one trained on male speakers and one on female speakers. The sex of each speaker determines which set of phone variants is specified in the supervision grammar

Each conversation is time-aligned by a hierarchical-grammar speech recognition algorithm, using the corresponding conversation, word, and phone models. This algorithm is similar to that described in [6], although more sophisticated in its treatment of multilayer grammars; the simplest evaluation metric, pooled covariance, is used. The recognizer outputs the beginning time and duration for each word. This output is then combined with the original transcription to produce a time-aligned transcription showing speaker turns. An excerpt is shown in Figure 2.

Speaker	Start	Duration	Word
B	68.66	0.34	Another
B	69.02	0.28	place
B	69.30	0.08	that
B	69.38	0.14	I
B	69.52	0.28	heard
B	69.80	0.14	is
B	69.94	0.30	really
B	70.24	0.46	pretty
B	70.76	0.34	is,
B	71.10	0.22	uh,
B	71.36	0.18	the
B	71.54	0.30	Cayman
B	71.84	0.26	Islands.
A	72.16	0.24	Oh
A	72.40	0.38	yeah.
B	73.02	0.26	Although
B	73.28	0.16	I
B	73.44	0.14	hear
B	73.58	0.10	they're
B	73.68	0.54	expensive.
A	74.22	0.24	Yeah,
A	74.46	0.18	that's
A	74.64	0.54	definitely
A	75.18	0.20	true.

Figure 2. Excerpt from time-marked transcript showing speaker, start time, and duration for each word.

Two interrelated issues that arose in defining this procedure are the use of the combined-channel signal versus the two single-channel signals, and the treatment of simultaneous speech. For processing efficiency, we preferred to align the combined-channel signal, since aligning each channel separately would require twice the processing time. In addition, alignment of the single-channel signal is vulnerable to errors associated with the "silent" portions of each signal, i.e., the times when the other participant was speaking. For example, some conversations contain considerable cross-channel echo, resulting in a relatively strong speech signal not reflected in a supervision grammar representing only one side of the conversation. This unrepresented signal tends to introduce spurious alignments, resulting in overall alignment failure.

The alternative approach, aligning the entire conversation with the combined-channel signal, requires an effective method of handling simultaneous speech segments. Stretches of simultaneous speech are labeled as such during transcription, but it is not generally feasible to specify a precise interleaving of words during simultaneous speech. Hence, a simple nonbranching supervision

B: Well, nice talking to you.
A: #All right.#
B: #Bye, bye.#
A: Bye.

Figure 3a. Transcription excerpt. Words delimited with # were spoken simultaneously.

```
nf_1034 —> WELL_f, nf_1035.
nf_1034 —> <BACKGROUND>, nf_1034.
nf_1035 —> NICE_f, nf_1036.
nf_1035 —> <BACKGROUND>, nf_1035.
nf_1036 —> TALKING_f, nf_1037.
nf_1036 —> <BACKGROUND>, nf_1036.
nf_1037 —> TO_f, nf_1038.
nf_1037 —> <BACKGROUND>, nf_1037.
nf_1038 —> YOU_f, nf_1039.
nf_1038 —> <BACKGROUND>, nf_1038.
nf_1039 —> ALL_m, nf_1040.
nf_1039 —> BYE_f, nf_1041.
nf_1039 —> <BACKGROUND>, nf_1039.
nf_1040 —> RIGHT_m, nf_1042.
nf_1040 —> <BACKGROUND>, nf_1040.
nf_1041 —> BYE_f, nf_1042.
nf_1041 —> <BACKGROUND>, nf_1041.
nf_1042 —> BYE_m, nf_0.
nf_1042 —> <BACKGROUND>, nf_1042.
```

Figure 3b. Corresponding conversation-level grammar rules. Branches allow either "all right" or "bye bye" to follow "you," but not both. Note that speaker A is male and speaker B is female.

grammar based directly on the transcription will not yield satisfactory alignment performance.

Our approach is to insert alternate paths in the grammar for the duration of the simultaneous speech portion. An example is shown in Figure 3. Constrained by such a grammar, the recognizer aligns the words for one participant or the other, but not both; it automatically selects between the two paths, based on which aligns better. This method has proven highly successful in enabling the alignment procedure to handle simultaneous speech without going astray. The disadvantage is that it yields word-level timing data for only one participant during simultaneous speech segments. However, since simultaneous speech is typically rather brief, even the unaligned words are localized to a small segment of the file.

3. RESULTS

The automatic time alignment procedure has been applied to 212 conversations recorded and transcribed as part of the Switchboard corpus. Alignment is executed on a Sun SPARCstation 2 with 64 Mbytes of memory and averages less than 5 times real time for each file.

# conversations	212	
# alignment failures	2	(0.9%)
ave. words per conversation	1518	
ave. turns per conversation	120	

Table 1. Alignment performance. Failures are instances where the recognition algorithm was unable to align the signal with the complete transcription.

For these 212 files, the average number of words per conversation is 1518, and the average number of turns is 120. These conversations manifest the variety of challenging characteristics typical of long-distance telephone speech (limited bandwidth, handset variation, channel noise, background noise), spontaneous speech (hesitation phenomena, misarticulations and mispronunciations, non-speech sounds), and multiple-channel interaction (simultaneous speech, channel synchronization variation, cross-channel echo). Participants represent a variety of American English accents and range in age from 20 to 60.

Table 1 shows the alignment performance obtained on these 212 conversations. The recognition algorithm failed to align the entire transcription with the signal in only two cases, yielding a failure rate of 0.9%. In one of these, the failure was due to severe signal distortion introduced by one participant's telephone.

For the 210 conversations which were fully aligned, the accuracy of the resulting alignments was evaluated auditorily. Quantitative measures of the alignment accuracy rate such as can be computed for the TIMIT database require an independently-determined time alignment as a standard for comparison, which is not available for this corpus. Instead, these files were audited in 5-second segments and instances where alignment word times deviated significantly from the observed time were noted. The overwhelming majority of words are correct at least to within a second; in fact, random checks of individual words indicate that the alignment is normally correct to within one or two 20-ms frames. The dominant error, with occasional misalignments of up to 2 seconds, involves hesitation sounds such as "uh." In the alignment procedure, these sounds are treated simply as normal lexical items, although they may in fact be prolonged beyond the duration typical of phonologically comparable words. A planned refinement of this method will provide trained recognition models for such sounds to accommodate their exceptional duration characteristics.

The time-marking procedure has proven to be robust to the variety of difficult conditions enumerated above. For example, alignment is not affected by the presence of non-speech sounds such as breath noises and laughter. Perfor-

mance is also satisfactory on simultaneous speech segments, with the recognizer selecting one participant's speech for alignment. The importance of special treatment for simultaneous speech is corroborated by the fact that we have observed alignment failures in cases where the transcriptionist failed to note simultaneous speech. Correction of the transcription in these cases enables successful automatic alignment.

The sole condition which has introduced significant failure rates is corruption of the digital signal induced by a T1 telephone interface hardware failure. In a set of 110 test files manifesting at least one instance of signal corruption, the alignment failure rate was 24% (26 failures). Failure was more common on files rated as significantly corrupted, i.e., where portions of the speech were unintelligible to transcribers. (Because of the signal corruption, these files are not designated for inclusion in the Switchboard corpus.)

Another challenge is posed by conversations in which the recording began at slightly different times for the two sides of the conversation. When the channels are combined, this asynchrony causes a noticeable lag between the signal and its cross-channel echo, which can be quite distracting when the echo is strong. The time alignment procedure has proven very robust to this condition, making it unnecessary to resynchronize the channels before time-marking the data. In fact, because the procedure remains reliable under such conditions, timing information can be used in analyzing and resynchronizing the data. For example, we can determine the degree of signal-echo separation in files containing perceptible echo on both channels, using the time alignment to identify echo portions of the file, as illustrated in Figure 4. By identifying the echo, this procedure also facilitates resynchronization based on cross-correlations, or echo cancellation if desired as a data preprocessing operation.

Thus, this method provides reliable time alignment with a relatively low expenditure of manual and computational effort. It is robust to a variety of anomalies and artifacts in the data, making it suitable for processing large amounts of speech collected without human supervision under the real-world conditions of the long-distance telephone environment.

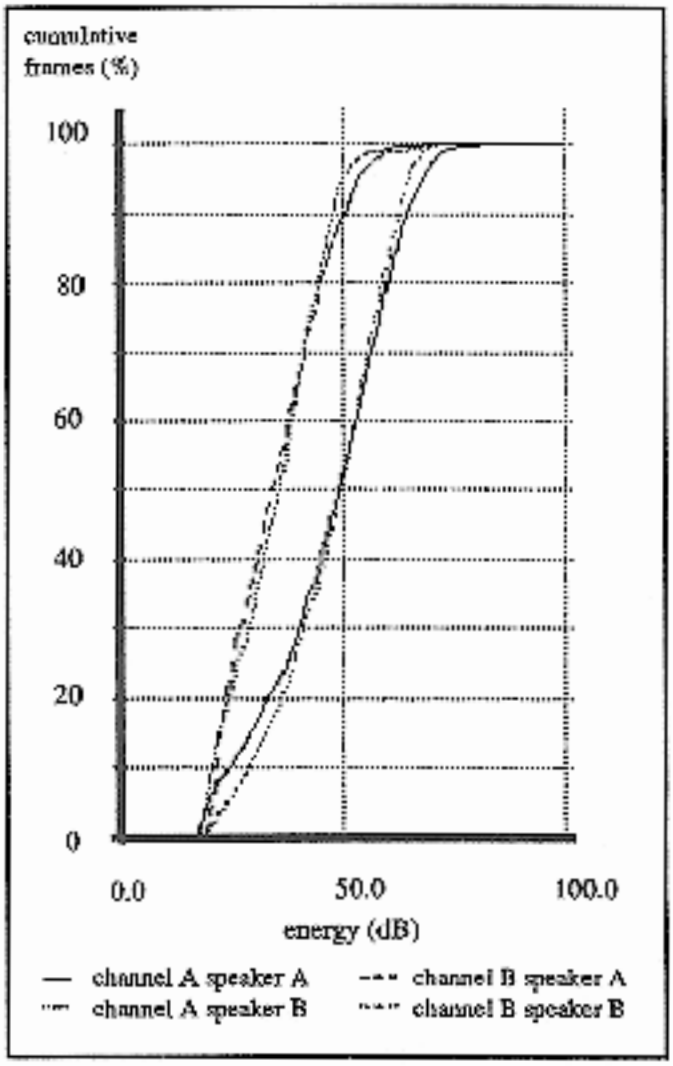

Figure 4. Signal-Echo separation in a file exhibiting strong echo in both channels. The time alignment of speaker turns was used to identify the echo.

REFERENCES

[1] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. ICASSP 92*, 1992.

[2] H.C. Leung and V. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP 84*, 1984.

[3] A. Ljolje and M.D. Riley, "Automatic Segmentation and Labeling of Speech," *Proc. ICASSP 91*, 1991.

[4] W. Fisher, V. Zue, J. Bernstein, and D. Pallett, "An Acoustic-Phonetic Database," *J. Acoust. Soc. Amer. Suppl. (A)* 81, 1987.

[5] B. Wheatley and J. Picone, "Voice Across America: Toward Robust Speaker-Independent Speech Recognition for Telecommunications Applications," *Digital Signal Processing* 1:2, 1991.

[6] G.R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," *Proc ICASSP 89*, 1989.