Aligning phonetic transcriptions with their citation forms

Keith Johnson
Department of Linguistics, Ohio State University, Columbus, OH 43210-1298
kjohnson@ling.ohio-state.edu

Abstract: One of the main motivations for publishing this paper is to make available a matrix of phone-distance measures which may be useful in dealing with large corpora of conversational speech. The paper reports how this matrix of phone-distances was created from transcriber labeling disagreements, and how it can be used in a dynamic time warping algorithm to align phonetic transcriptions of conversational speech with their citation forms. The weighted string edit distance produced by the phone-distance DTW algorithm may also be useful in calculating neighborhood densities for studies of auditory word recognition.

PACS numbers: 43.72.Lc, 43.72.Ne

1. Introduction

For phonetic and phonological analyses of speech corpora it is necessary to map phonetic transcriptions of actually produced speech onto citation forms of the words spoken, because the goals of many analyses involve tabulating patterns in the relationship between the citation forms of words and their realizations in speech. The patterns of interest in these types of analyses are substitutions, deletions, and insertions.

For example, in segmental descriptions of casual speech processes we say that *lean bacon* may be pronounced *lea[m] bacon* a substitution of [m] for [n], or *just now* may be *jus' now* a deletion of [t], or *jiust now* an insertion of [i]. To align the phonetic transcriptions of the citation pronunciation and of the actual pronunciation we need to perform a mapping that will correctly detect deletions and insertions and map similar segments onto each other.

Other areas of spoken language research also rely on string edit distance as a measure of word similarity. In psycholinguistic studies of auditory word recognition it has been found that lexical neighborhood density is an important parameter (Luce and Pisoni, 1998). Lexical neighbors are often determined by an edit distance rule attributed to Greenberg and Jenkins (1964) in which the number of insertions, deletions, and substitutions required to change one phonetic string into another is used as the word similarity metric. Words that are similar to a target word with a distance of one (i.e., they differ by a single substitution, insertion, or deletion) are considered to be lexical neighbors of the target word. The roughness of this measure is apparent in the neighbors of *reed* which include *lead*, *seed*, *keyed*, etc. where *lead* is more similar to *reed* than is *keyed*. Thus, a more sensitive measure of word similarity might be useful in psycholinguistic research.

This paper reports a variant of dynamic time warping (Wagner and Fischer, 1974; Sankoff and Kruskal, 1983) that is adapted for use with phonetically transcribed spoken language corpora. The goals of the algorithm were to provide an accurate, automatic, and consistent alignment of phonetic transcriptions onto citation forms. Some example alignments drawn from the "Variation in Conversation" (ViC) corpus (Pitt *et al*., 2003) are shown in table 1. Note the the phonetic forms are written in an extension of the ARPABET in which [iyn] stands for nasalized [iy], [em] is a syllabic [m], and [er] is a syllabic [r] [see Ohio State University (2002) for an explanation of the symbol set].

In one instance of the word *twiddling* the speaker deleted two segments [d] and [ng] and substituted [ah] for [ih] and [iyn] for [ih]. In *something* an initial [t] was inserted, the

[m] became a [p], the [th] and [ih] were deleted, and the final [ng] was replaced by [em]. In *predominantly* we see five deletions in a row in this three-syllable pronunciation of a five syllable word. In *hundred* four of the seven segments were deleted. "Massive" reductions like these are quite common in the corpus (Johnson, 2003) and thus a robust mapping algorithm was needed because pronunciations such as these with quite a substantial amount of deviation between the citation form and the actual pronunciation are difficult cases for a phonetic string mapping algorithm.

Table 1. Examples of mapping actual pronunciation onto citation form pronunciation. The term *ed* refers to the edit distance calculated in the dynamic time warping routine.

Word	*ed*	Citation form	Actual pronunciation
twiddling	0.527	t w ih d el ih ng	t w ah . l iyn .
something	0.672	. s ah m th ih ng	t s ah p . . em
predominantly	0.516	p r ih d aa m ih n ax n t l iy	p er . d aa m . . . . . l ax
hundred	0.669	hh ah n d r ax d	. ah nx . er . .

2. The ViC corpus

The dynamic time warping algorithm described in this paper was developed for use in analyzing variation in a large corpus, and was developed using phone similarity data produced in a transcriber reliability study done as a part of this corpus project. The Variation in Conversation (ViC) corpus is a large database of recorded conversational speech. The following description of the corpus is a brief synopsis of the fuller account given in Pitt et al. (2003).

Forty talkers were from the Columbus, OH community. All were natives of Central Ohio, and the sample was stratified for age (under 30 and over 40) and sex, and the sampling was limited to middle-class caucasians. Talkers were invited to come to the Ohio State University campus to have a conversation about everyday topics such as politics, sports, traffic, schools. After the interview, talkers were debriefed on the conversation's true purpose and all consented to having their speech used in research. Interviews were conducted in a small seminar room by one of two interviewers (one male and one female) who had been trained to conduct sociolinguistic interviews. Talkers sat in a chair facing the interviewer and wore a head-mounted microphone which fed into a DAT recorder.

Talkers spoke a total of 306,652 word tokens, of which ~100,000 have been phonetically transcribed with hand-corrected phonetic labels (this includes hesitation noises like *um* and *er*, which were not included in other studies of the same set of speakers, e.g., Johnson (2003). The size of the hand-labeled corpus is approximately twice the size of the TIMIT read-speech corpus (Zue et al., 1990). Phonetic transcription proceeds in three steps. First, an orthographic transcription is produced. Second, an HMM-based recognizer performs a forced alignment of dictionary pronunciations onto the acoustic signal (Wightman and Talkin, 1997). Third, a team of phoneticians (graduate students and post-docs in linguistics) hand-correct the aligner output. So far in our phonetic transcription effort, recordings of 14 of the 40 speakers have been phonetically tagged and served as the testbed for the DTW algorithm reported here.

3. Transcriber disagreements

A transcription consistency study was conducted using data from the ViC corpus (Raymond, 2003) and because the transcriber disagreement data from that study plays a central role in the phone-distance mapping algorithm, a brief description is in order. For a fuller description of the transcriber reliability study, see Raymond (2003). Four transcribers phonetically transcribed four 1-min samples from four different talkers following the conventions

documented in the project coding protocol (Kiesling and Raymond 2000). The speech samples started approximately 10 min into the interviews with four different talkers (young male, young female, old male, old female) and in total consisted of about 730 words (including dictionary words and other types of "word events", such as lexical cutoffs and fillers) and about 2300 phonetic segments (with four transcribers we have 13,800 pairs of phonetic transcriptions). Transcribers worked independently, starting with the extant English text transcriptions for each sample. The phonetic transcriptions were equated for comparison across transcribers according to time-stamps associated with the phonetic symbols to assure that comparisons were of phonetic labels that had been applied to the same stretch of speech.

Of prime interest here is that this procedure produced a transcriber disagreement matrix. For each pair of symbols in the phonetic alphabet we have a measure of how often one transcriber chose symbol 1 while another transcriber chose symbol 2. This matrix of transcriber disagreements provides a measure of the subjective similarity of the phonetic symbols to each other, and thus can be used to give a weight to each possible phone substitution in a dynamic time warping algorithm.

4. From disagreements to distance

For any two phones in the transcriber disagreement matrix we have a submatrix of four cells. So, for the phones *i* and *j*, we have a submatrix with the probability of transcriber disagreements when transcriber 1 chose symbol *i* and transcriber 2 chose symbol *j* or *vice versa* (p_{ji} *or* p_{ij}), and the probabilities of transcriber agreements when both transcribers chose the same symbol (p_{jj} *or* p_{ii}).

Shepard (1972) suggested a simple heuristic method for calculating psychological similarity and distance from such a matrix. Similarity, according to Shepard, is found by scaling the disagreements involving the two sounds by the agreements. Thus, Eq. (1) gives us a value S_{ij} which is the similarity between phone *i* and phone *j*. A small constant was added to the numerator in (1) to avoid $S_{ij} = 0$ in (2). Distance is then the negative of the natural log of the similarity [Eq. (2)]. This is Shepard's law, which states that the relationship between perceptual distance and perceptual similarity follows an exponential function:

$$S_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}} \quad , \tag{1}$$

$$d_{ij} = -\ln(S_{ij}) \quad . \tag{2}$$

I used this method with the Raymond (2003) transcriber disagreement matrix to calculate the distance between each phone in the ViC phonetic symbol set, producing a "phone distance" matrix. Because substitutions normally add a weight of 2 to the string edit distance in dynamic time warping, I scaled the distance values d_{ij} so that the maximum distance is 2 and all other distances between phones fall between 0 and 2. There are many instances of 2 in the phone distance matrix because many of the phones were never substituted for each other. For example, there was never a case where one transcriber chose to call a segment [t] and another transcriber labeled the same segment [aa], so the distance between [t] and [aa] is 2. On the other hand, segments that were transcribed [aa] by one transcriber were frequently transcribed [ao] by one of the other transcribers, so the calculated distance between these two phones turned out to be 0.623.

Some "touchups" of the phone distance matrix were required. For example, some symbols were not used in the transcriber reliability study, or occurred so infrequently as to provide faulty estimates of phonetic similarity. In these cases (3.5% of the total number of

symbol pairs) I estimated phonetic similarity using my best guess based on the calculated similarity of comparable pairs. For example, the nasalized variants of some but not all vowels were available in the reliability corpus. Nasalized vowels for which data were not available were given similarity values comparable to the nasalized vowels that did appear in the matrix (similarity in relation to their non-nasal counterparts as well as similarity in relation to nasal consonants). My subjectively estimated similarity values are listed in the confusion matrix with a single digit after the decimal point, while the automatically generated similarity values appear in the matrix with many digits after the decimal point.

The phone-distance matrix described in this section is available with this paper (Mm.1. Phone-distance matrix (20 Kb)).

5. Dynamic time-warping

In the simplest case (Sankoff and Kruskal, 1983; Wagner and Fischer, 1974), string edit distance is calculated by finding the best mapping between two strings using a weighting function such that a substitution of one symbol for another costs 2 (one deletion and one insertion), each deletion or insertion costs 1, and the identity mapping costs 0. So, for example the string edit distance between [s t aa p] *stop* and [t aa p] *top* is 1 ([s] deletion), the distance between [t ae p] *tap* and [k ae p] *cap* is 2 (substitution of [k] for [t]), and the distance between [t ay p] *type* and [ay] *I* is also 2 (deletion of [t] and [p]).

In this method, edit distance is only very roughly correlated with subjective distance or perceptual confusions. Consider for example the pairs *lead/reed* ([l iy d]/[r iy d]) and *lead/seed* ([l iy d]/[s iy d]). The standard string edit distance for these pairs does not differentiate them. They both are related by a substitution and thus have an edit distance of 2, which when normalized by the number of symbols in the pair (6) is 2/6 = 0.3333. On the other hand, string edit distance using the phone substitution distances calculated from transcriber disagreements (Sec. 4) does differentiate these pairs. The edit distance for *lead/reed* is 0.27 while the edit distance for *lead/seed* is 0.33. Thus, weighting the cost of a substitution by a measure of the phonetic distance between the symbols gives a better estimate of the apparent distance between the word forms and so may provide a better estimate of lexical neighborhood density.

Calculating string edit distance in this way should also provide a better mapping between phonetic transcriptions and their citation forms. I turn now to an evaluation of this claim.

6. Evaluation of the algorithm

To evaluate the phone-distance-weighted algorithm I compared alignments done using the phone-distance matrix to alignments done with the standard string edit distance DTW algorithm.

The test set is composed of the phonetically transcribed words in the ViC corpus, as described in Sec. 2 above. In total, the mapping between phonetic transcriptions and citation forms was calculated for 99,677 words using the standard string edit algorithm and the phone-distance-weighted algorithm. Usually the two methods gave the same mapping of phonetic transcription onto citation form, but for 8.4 % of the words (8347 productions) the two algorithms produced different mappings.

The key difference between the two algorithms is that the standard algorithm sometimes produces more than one “best” path relating the two strings whereas the phone-distance algorithm generally produces only one best mapping. The behavior of the standard algorithm is underspecified with respect to ties in string edit distance, so I implemented it to prefer substitutions over deletions and insertions. This causes the standard algorithm to map [k ae t] onto [k ae p] (with a substitution) rather than mapping [k ae . t] onto [k ae p .] (with one insertion and one deletion) even though the string edit distance is 2 in both of these mappings. This implementation of the standard algorithm produces the best match to the

phone-distance algorithm.

However, even with this optimal implementation, when the standard algorithm is faced with both deletions and substitutions in a mapping, it stacks up the deletions at the beginning of the word, while the phone-distance algorithm lines up similar, but nonidentical, phonetic elements in the two strings. The improvement offered by the phone-distance algorithm is evident in examples in which the two algorithms produce different mappings between citation form and phonetic transcription (see Table 2). The examples in Table 2 are typical cases that illustrate the general pattern of behavior seen in most of the 8347 cases in which the two methods differ.

Consider for example the word *that*. The phone-distance algorithm correctly lined up the vowels [ae] and [ax], showing that the final [t] was deleted, while the standard method placed the deletion early in the string suggesting that [ax] is a substitution for [t]. The early placement of deletions is also apparent in longer words. For example, in one instance of *forgot* the standard method lines up the syllabic [er] with [g] while the phone-distance method correctly aligns [r] and [er].

Table 2. Examples of the mapping from phonetic transcriptions to citation forms produced by the phone-distance weighted and the standard DTW algorithms.

Word	Phone-distance	Standard
that	dh ae t dh ax .	dh ae t dh . ax
forgot	f ow r g aa t f . er . aa t	f ow r g aa t f . . er aa t
pregnant	p r eh g n ax n t p r eh g n axn . .	p r eh g n ax n t p r eh g . . n axn
material	m ax t ih r iy el m ax t iy r . el	m ax t ih r iy . el m ax t . . iy r el
probably	p r aa b ah b l iy p r ay	p r aa b ah b l iy p r . . . . . ay

In the production of *pregnant* shown in Table 2, the phone-distance algorithm's ability to line up [ax] with its nasalized counterpart [axn] shows that the final [n t] cluster of the citation form was deleted in this production. The analysis given by the standard algorithm is that the medial [n ax] were deleted and that [t] was realized as [axn].

The next example is another case where the standard algorithm's tendency to prefer deletions over substitutions leads to a misalignment that is avoided by the phone-distance model. The word *material* was pronounced with [iy] in the second syllable instead of [ih]. The standard algorithm posits deletion of [ih] and [r] in order to line up the transcribed [iy] with the citation form [iy]. Then it has to posit that an [r] was inserted. The phone-distance algorithm avoids this mistake by accepting the substitution of [iy] for [ih] in the second syllable.

There are cases in this database that indicate that the segmental analysis implicit in phonetic transcription is inadequate to account for the patterns of phonetic reduction in conversational speech. In some of these cases it seems as if the standard algorithm's alignment might be just as correct as the mapping given by the phone-distance algorithm. For example, in one production of *probably* we have the pronunciation [p r ay]. The diphthong in this production seems to be a coalescence of the first and last vowels of the

word (and who knows, maybe the [ah] vowel is in there too). The phone-distance mapping aligns [ay] with the first vowel [aa], while the standard algorithm puts off the substitution to the end of the string, thus lining [ay] up with [iy].

In conclusion, the mapping algorithm described in this paper, and crucially the use of an empirically derived matrix of phone distances, provides a better mapping between phonetic transcriptions and citation forms than can be produced by a standard dynamic time warping algorithm. These two algorithms give different mappings in 8.4% of the words in the ViC corpus, and in all of the cases examined for this study the phone-distance algorithm produced an equally good (in unusual cases like *probably*) or better mapping than did the standard DTW algorithm.

Acknowledgments

This work was supported by NIH Grant No. R01 DC04330–02. Mary Beckman and Chris Brew commented on an earlier version of this paper.

References and links

Greenberg, J.H. and Jenkins, J.J. (1964). "Studies in the psychological correlates of the sound system of American English," Word **20**, 157–177.

Johnson, K. (2003). "Massive reduction in conversational American English," in *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*. August, 2002, Tokyo; http://vic.ling.ohio-state.edu/massive_reduction.pdf

Kiesling, S. and Raymond, W.D. (2000). "The ViC transcriber's manual: Guidelines for transferring, transcribing, and labeling sound files for the *Buckeye* corpus." Ohio State University; http://vic.ling.ohio-state.edu/manual.html

Luce, P.A. and Pisoni, D.B. (1998). "Recognizing spoken words: The neighborhood activation model," Ear Hear. **19**, 1–36.

Ohio State University (2002). "Symbol set used in the ViC corpus," http://vic.ling.ohio-state.edu/alpha.html

Pitt, M., Johnson, K., Hume, E.V., Kiesling, S., and Raymond, W.D. (2003). "The ViC corpus of conversational speech," http://vic.ling.ohio-state.edu/IEEEdraft.pdf

Raymond, W.D. (2003). "An analysis of coding consistency in the transcription of spontaneous speech from the Buckeye corpus," in *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*. August, 2002, Tokyo; http://vic.ling.ohio-state.edu/SSDA.ms.pdf.

Sankoff, D. and Kruskal, J. (editors) (1983). *Time Warps, String Edits, and Macromolecules.* (Addison-Wesley, Reading, MA) reprinted (Center for the Study of Language and Information, Stanford, CA, 1999).

Shepard, R.N. (1972). "Psychological representation of speech," in *Human Communication: A Unified View*, edited by E.E. David and P.B. Denes. (McGraw-Hill, New York) pp. 67–113.

Wagner, R. and Fischer, M.J. (1974). "The string-to-string correction problem," J. Assoc. Comput. Mach. **21**, 168–173.

Wightman, C.W., and Talkin, D.T. (1997). "The aligner: Text-to-speech alignment using Markov models," in *Progress in Speech Synthesis*, edited by J. van Santen (Springer-Verlag, New York), pp. 313–323.

Zue, V., Seneff, S., and Glass, J. (1990). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**, 351–356.