ROBUST AUTOMATIC EXTRACTION OF DIPHONES WITH VARIABLE BOUNDARIES

Debra Yarrington, H. Timothy Bunnell, & Gene Ball*
email: yarringt@asel.udel.edu
Applied Science and Engineering Laboratories
University of Delaware/ A. I. duPont Institute
Wilmington, Delaware, USA 19899

*Microsoft Corporation
Redmond,Washington, USA 98052

ABSTRACT

This paper presents the approach under development at the Applied Science and Engineering Laboratories (ASEL) for automatic extraction of diphones from a speech database. The present system operates on a set of digitized spoken carrier words to (a) assign segment boundaries within the carrier words, (b) select the best instances of each diphone among the carrier words containing that diphone, and (c) assign multiple, context conditioned, boundaries within the selected diphones. Experiments designed to test the intelligibility and naturalness of the automatically extracted diphones indicated that the automatically extracted diphones resulted in synthesized speech that was slightly more natural sounding and slightly less intelligible than speech synthesized from manually extracted diphones of the same talker.

1. INTRODUCTION

One of the design objectives of ASEL's synthesizer is the ability to simply and rapidly change "voices" by changing diphone inventories. To this end, we are particularly interested in developing robust automatic diphone extraction software capable of building diphone inventories with minimal need for manual intervention by trained technicians.

2. ALGORITHM

The extraction process begins with a complete set of digitized carrier words spoken by a single talker. The present carrier word list comprises 1226 multisyllabic words. Each word is spoken 2 times by the talker. It is from this carrier word set that diphones are selected for the inventory.

2.1 Segmentation

Segment boundaries are assigned within each of the carrier words by a template matching process. In this process, a synthetic standard for each carrier word is synthesized using an existing diphone inventory and hence, diphone boundaries within the synthetic standard are known exactly. Acoustic analyses of both the spoken carrier word and the synthetic standard provide feature vectors for each 10-msec frame of speech. The feature vectors consist of the first 6 Bark-cepstrum coefficients. A dynamic time warping algorithm is then used to obtain the best alignment between the feature frames of the spoken word and the synthesized word. Segment boundaries within each carrier word are then heuristically derived from assigned diphone boundaries.

2.2 Diphone Selection

Following segment boundary assignment, the best instance of each diphone is selected from the carrier set. This is accomplished through a 2-pass process. In the first pass, up to five instances of each diphone are chosen. These diphones are chosen based on the comparison of their two parent phonemes with all other examples of those phonemes in the carrier word set. Each phoneme is assigned a rating based on its distance in the feature space from every other like phoneme. Specifically, this distance D between two instances of a phoneme is determined as:

$$D_{ij} = \mathrm{MIN}\,(d) \left\{ \begin{array}{l} \text{for } i = 1, I \\ \quad \text{for } j = 1, J \\ \qquad d = \sqrt{\displaystyle\sum_{K=1}^{M} \left(V_{ik} - V_{jk}\right)^2} \end{array} \right. \tag{1}$$

where I and J are the number of frames in the I^{th} and J^{th} instances of the phoneme, and M is the number of coefficients in each feature vector. A goodness rating is then assigned each phoneme as the total of its pair-wise distances. That is:

$$R_p = \sum_{I=1}^{N-1} D_i \tag{2}$$

where N is the number of instances of phoneme p in the carrier word set. Based on these phoneme ratings, up to five examples of each *diphone* are chosen by selecting those whose parent phonemes are lowest ranked. These selected diphones then function as the reduced word list for the second pass of the selection procedure. This second pass is otherwise identical to the first pass. This procedure results in a final set of diphones that is set-wise most consistent.

2.3 Boundary Set Selection

The final step in the current automatic extraction process is the assignment of context conditioned boundaries within each of the selected diphones. Rather than assign a single, and probably suboptimal, boundary at the edge of each diphone, a set of boundaries define the edges of each diphone, with the actual boundaries used in concatenation conditioned by the diphone context. To determine these boundaries, both spectral and temporal factors are considered. Specifically the distance between 2 phoneme vectors is calculated as the Euclidean distance between the 2 spectral feature vectors multiplied by a duration weighting factor. The weighting factor is determined using the formula:

$$w = \begin{cases} (2-x)^2 & \text{if } x < 1 \\ x^2 & \text{else} \end{cases} \quad (3)$$

$$x = \frac{t_1}{d_1} + \frac{t_2}{d_2} \quad (4)$$

where t_1 and t_2 are the frame locations in milliseconds for phoneme 1 and phoneme 2, respectively, and d_1 and d_2 are the phoneme durations in milliseconds of phoneme 1 and phoneme 2 respectively. The weighting factor (a) aids in keeping the duration of the synthesized phonemes close to their natural durations; and (b) reduces the chances of plosives being optimally matched at the silences such that no burst is included.

The location of the 10-ms frames with the best spectral match between 2 phonemes is then refined to avoid waveform discontinuity in the time domain. This is a matter of moving the exact boundary locations to the positive going zero crossing at the onset of a pitch period (for voiced speech) or to the first positive going zero crossing in the 10-ms frame (for voiceless speech). Diphones are then extracted from their carrier words at their most encompassing boundaries.

3. PERCEPTION EXPERIMENT

To objectively judge the quality of the automatically extracted diphones, a study was designed to compare the intelligibility and naturalness of speech synthesized with diphones from the automatic extractor to speech synthesized with diphones extracted manually. Two sets of 160 nonsense words were synthesized. One set was synthesized from the diphones created by the automatic extractor (the AUTOMATIC condition). The other set was synthesized from a set of diphones that had been previously extracted manually from the same set of carrier words (the MANUAL condition).

3.1 Stimuli

Most words chosen for synthesis were from a corpus designed to test intelligibility of consonants and consonant clusters in synthetic speech (Spiegel, *et al.*, 1990). Nine of the 158 words in this Bellcore list were not included in the present study because the necessary diphones were unavailable. Sixty-two of the words were designed to test word-initial consonant(s) (ONSET group), and 87 were designed to test word-final consonant(s) (OFFSET group). Eleven other words were added to test nuclear vowel sounds (VOWEL group).

3.2 Subjects

The subjects were 22 young adults (16 males and 6 females) who were all native speakers of American English. All subjects reported having normal hearing.

3.3 Procedure

Subjects were instructed to type an English spelling representation of what they heard. They were also told to rate the naturalness of each word from 1 to 5, with one being very synthetic sounding, and 5 sounding as if a human had said it. The subjects listened to a practice session of 15 words. Upon successful completion of the practice session, subjects listened to all synthetic stimuli, blocked by diphone set (MANUAL or AUTOMATIC). Half of the subjects heard the AUTOMATIC block first, and half heard the MANUAL block first. Within each block, a different randomization was used for each subject.

3.4 Scoring

The responses were collapsed over subjects and stimulus conditions and every unique response was identified. Each unique response was then classified based on the parts of the word that were identified correctly (i.e., ONSET, VOWEL, and OFFSET). This classification was then used as a key to score the individual subject responses. This provided four dependent measures that could be used in subsequent analyses: ONSET, OFFSET, VOWEL, and WORD correct.

3.5 Data Analysis

Each of the dependent measures was used in separate T-tests comparing responses from the AUTOMATIC and MANUAL sets. These analyses were performed using subjects as the sampling units (scores were based on percentages of ONSET, OFFSET, etc. correct, or averaged naturalness ratings) and also using words as the sampling units (scores were percentages of subjects who responded correctly for ONSET, OFFSET, etc.). Separate T-tests were also run for the complete 160 word set and for the subgroups intended only to test ONSET, OFFSET, etc. Because the pattern of results did not differ greatly for any of the intelligibility measures, we report WORD correct scores for the entire group of 160 words only.

4. RESULTS

For analysis by subject, the results of the T-test indicated that percentage correct WORD scores were significantly better for the MANUAL stimuli than for the AUTOMATIC stimuli with $T = -3.15$, $p < 0.01$. However, the MANUAL stimuli were rated significantly less natural sounding than the AUTOMATIC stimuli with $T = 3.45$, $p < 0.01$.

For the by-word analysis, the naturalness ratings were again significantly higher for the AUTOMATIC stimuli with $T = 2.32$, $p = 0.02$. In this analysis the difference in mean percentage correct WORD scores was not significantl ($T = -1.61$, $p = 0.11$).

The means underlying these effects are shown in Figure 1 which plots the percentage correct WORD scores for both MANUAL and AUTOMATIC sets (left axis scale) and average naturalness ratings for these sets (right axis scale).

5. DISCUSSION

The results of this study are encouraging. The intelligibility of speech synthesized from the automatically extracted diphones was consistently close to that of the manually extracted diphones, and was consistently rated as more natural sounding than the manually extracted diphones. These results were accomplished with no manual intervention on behalf of the automatically extracted diphones. In actual use, efforts should be made to find and correct poorly extracted diphones. Our MANUAL set in this experiment had the benefit of such

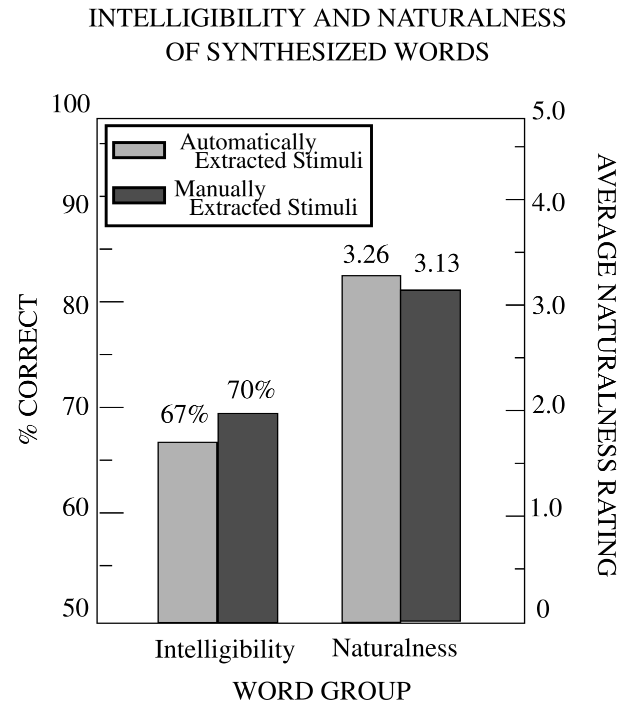

Figure 1: Intelligibility and naturalness ratings of stimuli from automatically and manually extracted diphones.

refinement even though our AUTOMATIC set did not. An important question pertaining to refining automatically extracted diphone inventories is whether a few bad diphones were responsible for the lowered intelligibility of the AUTOMATIC set, or whether the diphones of the AUTOMATIC set were consistently slightly poorer in quality than those of the MANUAL set. If the first was the case, it should not be too difficult to find and correct the bad diphones. The second case, however, is more insidious. Not only would errors be more prevalent, but errors would also be subtler and thus more difficult to find and correct.

If a few bad diphones were responsible for the lower intelligibility, we should find that certain automatically extracted stimuli were consistently misunderstood by subjects. Some support for this hypothesis can be seen in Figure 2 which plots the number of subjects who identified specific words correctly against the number of words so identified for both AUTOMATIC and MANUAL sets. For instance, looking at the left side of the figure, seven words in the AUTOMATIC set were never correctly identified, whereas, from the right side of the figure, there were 32 words from the AUTOMATIC set and 19 words form the MANUAL set that were correctly identified by 21 or more subjects. There were a total of 11 words in the AUTOMATIC set that were identified correctly by fewer than 2 of the subjects, whereas none of the words in the MANUAL set were this pooly recognized. Thus it is not the case that all words in the AUTOMATIC set tended to be a little less intelligible than words in the MANUAL set.

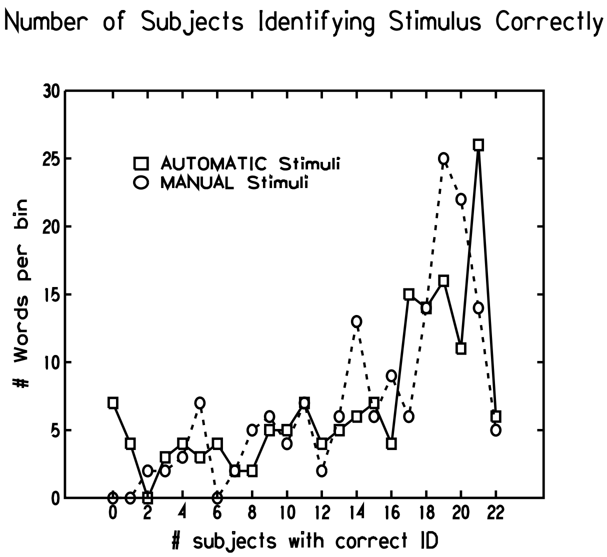

Figure 2: Words grouped by count of number of correct identifications by subjects.

One reason for possible failure in diphone extraction is a poor initial labeling of the carrier word. When the labeler failed, it tended to place all segment labels in the initial few milliseconds of silence before the start of the word. If more than one instance of a phoneme was mislabeled within the initial silence, these instances would match extremely well because silence matches extremely well with silence. Thus, those phonemes would receive an extremely high rating. A diphone whose parent phonemes had such high ratings would have a good chance of being chosen as the best example of that diphone.

The labeler could also fail by mislabeling 2 phonemes with the same neighboring phoneme. If these 2 instances were both parents of diphones chosen for the final set of diphones, their optimal spectral match could be in the adjacent neighboring phoneme. It is possible that the particularly poor tokens in the present experiment resulted from initial segmentation failure of these two types.

Regarding the naturalness ratings, we can speculate that speech from the AUTOMATIC set sounded more natural because diphones were appended at locations that were more spectrally similar than for the MANUAL set, thus resulting in smoother, more flowing speech with fewer perceptually salient discontinuities caused by large spectral mismatches. It is also possible that the automatically extracted diphones resulted in more natural sounding speech because syllable duration was better preserved. That is, the duration weighting factor had some influence on the boundary locations selected so that when speech was synthesized, each phoneme duration could be synthesized closer to its length in natural speech context.

6. CONCLUSION

The approach presented in this paper seems promising. Our approach resulted in speech that was consistently rated more natural and was consistently close in intelligibility to speech synthesized from diphones that were manually extracted from the same set of carrier words. An advantage our approach has over fixed boundary schemes (e.g. Kaeslin, 1986) is that spectral mismatches at diphone junctures are less likely to occur. Our approach extends existing variable boundary schemes (e.g. Conkie & Isard, 1994) by selecting setwise most consistent diphones including duration when determining individual boundary sets.

7. ACKNOWLEDGEMENT

This work is supported by The Rehabilitation Engineering Research Center on Augmentative and Alternative Communication (Grant #H133E30010-94) of the National Institute on Disability and Rehabilitation Research, U.S. Department of Education, the Nemours Research Programs, and the Microsoft Corporation.

8. REFERENCES

Conkie, A. and Isard, S., “Optimal Coupling of Diphones”, Conference Proceedings of 2nd ESCA/IEEE Workshop on Speech SYnthesis, pp. 119-122, New Paltz, New York, USA, 1994

Kaeslin, H. “A Systematic Approach to the Extraction of Diphone Elements from Natural Speech”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 2, April, 1986, pp. 264-271.

Spiegel, Murray F., Altom, Mary Jo, Macchi, Marian J., and Wallace, Karen L., “Comprehensive Assessment of the Telephone Intelligibility of Synthesized and Natural Speech”, Speech Communication, Vol. 9, No. 4, August 1990, pp. 279-291.