

Phonetic Effects on Listener Detection of Vowel Concatenation

Ann K. Syrdal

AT&T Labs - Research
Florham Park, NJ 07932-0971
syrdal@research.att.com

Abstract

Concatenative speech synthesis quality depends in part on the minimization of audible discontinuities between two successive concatenated units. This study focuses on human detection of concatenation discontinuities in synthetic speech. Statistical analyses compared for various phonetic categories the results observed in perceptual tests with two voices – one female and one male. Neither a comprehensive phonetic analysis nor a comparison of discontinuity detection between voices has been reported previously. Although discontinuities were generally more detectable for the female than the male voice, there were many similarities between results obtained from the two speakers. A reliably higher rate of detection of discontinuities was observed for diphthongs than for monophthong vowels. Post-vocalic consonants influenced concatenation discontinuities significantly more than pre-vocalic consonants, and post-vocalic sonorants were associated with higher detection rates than post-vocalic non-sonorants. The differences in discontinuity detection among vowels and among consonantal contexts for both voices consistently suggest that highly audible discontinuity is related to concatenation in regions of spectral change.

1. Introduction

Concatenative speech synthesis, in which small units of recorded speech are concatenated together to generate novel utterances, is an increasingly popular method for text-to-speech synthesis. Typically, diphone concatenative synthesis involves concatenation of diphone units in the approximate middle of a phone. A diphone is the last half of the first phone and the first half of the second phone in a sequence. A common and annoying artifact of concatenative synthesis is an audible discontinuity at concatenation join points[1].

Human detection of concatenation discontinuities in synthetic speech has rarely and only recently been studied. Only one other study [2][3][4] has directly focused on this issue. In that study, as well as the current study, listeners judged whether or not a synthesized test word contained an audible discontinuity. A second study [5] is somewhat related, but it focused instead on perceptual similarity judgments, which are more relevant to target costs as used in unit selection concatenative synthesis. Target costs are intended to estimate the similarity of a given unit in the speech database to an ideal unit. Concatenation costs estimate the detectability of concatenation discontinuities in unit selection synthesis. Both target costs and concatenation costs, when summed over all units that form a synthetic utterance, determine the sequence of units in a voice database that is optimal for that utterance. Due to the design of the concatenated test stimuli in [5], listeners' ratings would have been influenced by several factors, making interpretation of the results somewhat ambiguous with respect to concatenation discontinu-

ity detection: Using recorded CVC words, the second half of a vowel taken from a different consonantal context was spliced into another recorded CVC word. The synthetic test stimuli therefore had two concatenation points rather than one, and in addition, they contained potentially conflicting cues to several features of the final consonant (one set of cues from the preceding spliced vowel and another set from the final consonant itself – for a discussion of this topic, see [6]), which would also affect listeners' judgments of similarity between the original recorded word and the synthetic concatenated version.

Both previous studies focused on using human perceptual results to evaluate predictive algorithms based on spectral representations of units at the concatenation point. We have also reported a similar analysis with some of our perceptual data [7]. The current study includes a more comprehensive phonetic analysis of detection data from two voices: one female and one male. It is not known how concatenation detection results generalize across speakers, as no comparisons have been made previously.

A phonetic analysis of detection results is valuable because it indicates which acoustic features are associated with the audibility of concatenation discontinuities. Such knowledge is useful in improving algorithms or suggesting strategies to avoid perceptible concatenation discontinuities in speech synthesis.

The only phonetic analysis of concatenation discontinuity detection reported in the previous studies was limited to detection rates among a set of 5 Dutch vowels [2][3][4]. Vowels are important as an initial focus because their relatively higher energy makes concatenation discontinuities more salient.

Rigorous psychophysical signal detection[8] methods were used in the perceptual experiment. These include a large number of test utterances and the inclusion of ample false alarm conditions to provide for the calculation of d' , a standard psychophysical index of perceptual detectability. d' is essentially the normalized difference between the means of two distributions along a sensory dimension: in the case of the present experiment, one distribution is conditioned upon the presentation of speech with no discontinuity, and the other distribution is conditional upon speech with discontinuity. A high d' thus corresponds to high detectability of concatenation discontinuity.

2. Methods

A psychoacoustic experiment was conducted on listeners' detection of concatenation discontinuities in a large number of test words generated by concatenative synthesis using speech data from two speakers.

2.1. Test stimuli

A set of 2016 monosyllabic test words were generated by concatenative synthesis using an acoustic inventory of recordings from one adult female speaker, and an equal number were generated from one adult male speaker. An experimental version of the AT&T NextGen text-to-speech (TTS) synthesizer [9] was used to synthesize the test stimuli.

The acoustic inventory used for synthesis consisted entirely of recordings of the 336 monosyllabic test words that constitute the Modified Rhyme Test (MRT)[10][11][12], a standard test of speech intelligibility [13]. The MRT is composed of 56 sets of six similar words. The six words within a set differ by either the initial consonant(s) (such as “book, took, shook, cook, hook, look”), or by the final consonant(s) (such as “sing, sit, sin, sip, sick, sill”), and all words in a set contain the same vowel nucleus. The former group of 28 6-word sets will be referred to as the Initial Different/Final Same (IDFS) group, and the latter group of 28 6-word sets will be termed the Initial Same/Final Different (ISFD) group. In several instances, sets contain a word or words that are either vowel-initial (such as “oil, foil, coil, boil, toil, soil”) or vowel-final (such as “ray, raze, rate, race, rake, rave”). A restricted domain system for each voice was built with the MRT inventory using the NextGen TTS system.

For each of the two voices, 2016 synthetic test stimuli were synthesized by concatenation of selected portions of the 336 recorded words contained in the acoustic inventory. Each recorded word in the inventory was essentially divided into two parts, its initial and final halves. The initial half consisted of the word-initial consonant(s) (if any) and the first half of the vowel nucleus. The second half consisted of the second half of the vowel nucleus and the word-final consonant(s) (if any). From each 6-word set, 36 test stimuli were synthesized. All possible combinations of the 6 initial halves and 6 final halves within a set were concatenated to generate 36 synthetic test words. Of the 36 test words synthesized from each 6-word set, 30 combined the first half of a word with the second half of a different word, and these 30 test words had the potential of containing detectable concatenation discontinuities. Six of the 36 test words synthesized per set were resynthesized versions of the first and second halves of the same word, and they would be expected to contain no detectable concatenation discontinuities. This process was repeated for each of the 56 6-word sets, yielding 2016 test words.

A very simple concatenation method was used by the synthesizer to concatenate the first and second halves of words at approximately the mid-point of the vowel. Using the raw waveforms, the concatenation point was determined by a minimum in the cross-correlation function calculated over a narrow window around the vowel mid-points. In this way, concatenation discontinuities due simply to arbitrary abutment of the two halves was avoided, and pitch period continuity was maintained. The original fundamental frequencies of the two constituent word halves was unaltered.

2.2. Procedure

The female voice and male voice tests were conducted independently, but they followed the same procedure. The listening test followed a simple single interval forced choice Yes/No signal detection paradigm commonly used in psychoacoustic experiments. After hearing a test stimulus, a listener reported whether or not (s)he heard a concatenation discontinuity. Each stimulus was presented once per listener. The entire test battery

was divided into a series of subtests; each subtest contained 72 test stimuli and normally took under 10 minutes to complete. Each listener received a different randomization of the stimuli in a subtest. Typically, a listener would participate in no more than one subtest a day. Written instructions to listeners and one example of a stimulus for each response type (a detectable concatenation discontinuity and no discontinuity) were provided at the beginning of a subtest. Listeners were automatically prompted if they did not complete any part of the subtest, and their complete response record was stored in a log file identifiable by listener and subtest.

Listening tests were web-based and interactive. Listeners normally took the tests from workstations or PCs in their quiet private walled offices using the relatively high quality audio equipment normally available there. Listeners initiated the presentation of each stimulus by clicking an icon. Concatenation detection responses were made by clicking one of two button icons (one indicating that a discontinuity was detected, and the other, that no discontinuity was detected). Listeners were encouraged to use headphones, and the large majority indicated that they did so. The volume was adjusted to suit their individual preferences. Stimuli were sampled at 16 kHz.

2.3. Listeners

Eighteen adult volunteer listeners participated in at least one listening subtest: 16 during the test with the female voice and 12 for the male test. Ten served as listeners for both voices. All listeners were employees or contractors working at AT&T Labs Research. They represented diverse language backgrounds, since native language was not considered relevant for the auditory task of detecting concatenation discontinuities. The hit rate (correct detections), false alarm rate (false detections), and corresponding d' per subtest were monitored for each listener. Rarely (5% of the time for the female and 6% of the time for the male voice test), a listener's responses were rejected for a particular subtest if their d' score was substantially lower than the other listeners' d' scores for that subtest. There were at least five acceptable listeners for every stimulus word in the test set, and the average was 5.9 acceptable listeners per subtest with the female voice, and 7.75 per male voice subtest. There were 11,808 total acceptable observations in the female voice listening test, and 15,624 in the male test.

3. Results

Pooling all the acceptable listeners' responses for the female test, the hit rate was 61.4% and the false alarm rate was 6.1%. The group hit rate for the male voice was 54.2% and the false alarm rate was 7.4%. These results yield a d' score of 1.83 for the female voice, and 1.57 for the male voice, representing overall perceptual performance. Results for only the test stimuli for which the first and second halves were concatenated from different recorded words are presented in the remainder of the paper.

3.1. Overall analyses

Fig. 1 is a plot comparing overall detection (hit) rates, ISFD group detection rates, and IDFS group detection rates for the female and male voices. The higher overall detection rate evident for the female than for the male voice was statistically significant ($t = 6.0593$, $df = 1679$, $p < 0.0001$). There was a low but significant correlation between detection rates for female and male voices (Pearson's product-moment correlation

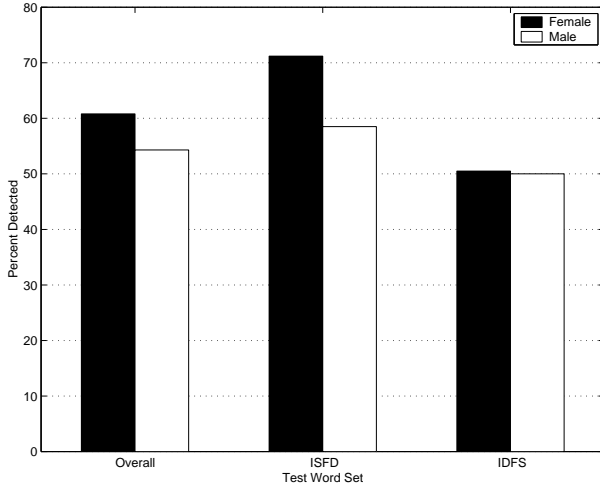


Figure 1: Detection rates for female and male voices.

$r = 0.27, t = 11.4844, df = 1678, p < 0.0001$.

3.1.1. Detection rates by vowel

Table 1 lists detection rates for each of 11 vowels for female and male voices (DARPAbet symbols are used to represent the vowels). There was no significant difference between the two distributions ($\chi^2 = 110, df = 100, p < 0.2322$). For each voice, detection rates for three classes of vowels were compared: diphthongs (/ey, ow, oy/), long monophthongs (/aa, ae, ao, iy/), and short monophthongs (/ah, eh, ih, uh/). For both voices, detection rates for diphthongs were significantly higher than for long vowels or for short vowels (Female Diphthong vs. Long Vowel: $t = 4.448, df = 833, p < 0.0001$; Female Diphthong vs. Short Vowel: $t = 6.6442, df = 1168, p < 0.0001$; Male Diphthong vs. Long Vowel: $t = 9.6265, df = 833, p < 0.0001$; Male Diphthong vs. Short Vowel: $t = 7.2564, df = 1168, p < 0.0001$). Comparing long and short monophthongs, the results were inconsistent; for the female voice, detection for long vowels was higher than for short vowels, but the opposite was true for the male voice. In [2][3][4], the Dutch vowel /a/ had

Vowel	Example	N Words	Female %	Male %
aa	cot	100	48	38
ae	cat	165	66	50
ah	cut	215	61	57
ao	caught	35	75	57
eh	pet	180	59	45
ey	Kate	240	72	66
ih	kit	420	55	56
iy	key	210	60	45
ow	coat	55	79	72
oy	coy	30	65	92
uh	cut	30	23	40

Table 1: Detection rates and number of test words by voice for 11 vowels

the lowest detection rate (17%) observed. In the current study, its American English counterpart /aa/ had considerably higher rates for both speakers, but nonetheless it had the second-lowest

detection rates observed. The lowest rates observed in the current study were for /uh/, which was not included among the five vowel Dutch set. Of the remaining three vowels included in both studies, the American English versions had slightly higher discontinuity detection rates for /iy/ and /ah/, and equivalent rates for /ih/. In the Dutch study, consonantal contexts were equated for all five vowels studied, however in the current study, they were not the same for each vowel. It is possible, therefore, that different consonantal contexts could have influenced detectability differences among the vowels.

3.2. Initial Same/Final Different (ISFD) test words

As is evident in Fig. 1, concatenation discontinuities are significantly more perceptible in ISFD test words than in IDFS words for both voices (Female: $t = 11.4835, df = 1678, p < 0.0001$; Male: $t = 5.0981, df = 1678, p < 0.0001$). For ISFD words, a significantly higher detection rate was observed for the female than for the male voice ($t = 9.1933, df = 839, p < 0.0001$). No strong or consistent effects of pre-vocalic consonant on detection of discontinuity were observed.

3.2.1. Effects of word endings on discontinuity detection in preceding vowel

For both speakers' ISFD test words, if the coda (final consonant or consonant cluster) of either of the two constituent words contained a sonorant (/l, r, m, n, ng/) or either constituent was vowel-final, discontinuity detection rates were significantly higher than if they did not (Female: $t = 16.0504, df = 838, p < 0.0001$; Male: $t = 9.9398, df = 838, p < 0.0001$). Table 2 lists for both voices the discontinuity detection rates for these two classes. For the sonorant coda set of ISFD test words, there

Class	N Words	Female %	Male %
Sonorant	476	86	68
Non-Sonorant	364	52	46

Table 2: Female and Male Concatenation Discontinuity Detection Rates and Number of Test Words for ISFD Test Words with Sonorant and Non-Sonorant Classes of Word Endings

was a significantly higher detection rate for the female than for the male voice ($t = 10.8628, df = 475, p < 0.0001$). For the non-sonorant set, the female voice also had a higher detection rate, but the difference between female and male rates was much smaller ($t = 2.3928, df = 363, p < 0.0172$).

3.3. Initial Different/Final Same (IDFS) test words

As illustrated in Fig. 1, for IDFS test words there was no significant overall difference between detection rates for female and male voices ($t = 0.2609, df = 839, p < 0.7943$). No large or consistent effects on detection rates were observed for various classes of onsets (word initial consonants or consonant clusters).

However, even though the codas of IDFS test word constituents were always the same, detection rates for IDFS words with sonorant codas were significantly higher than those with non-sonorant codas for both voices (Female: $t = 8.4208, df = 838, p < 0.0001$; Male: $t = 11.5763, df = 838, p < 0.0001$).

Detection rates for IDFS test words with sonorant codas did not differ between female and male voices ($t = 0.7485$,

$df = 449, p < 0.4545$), nor did detection rates for IDFS words with non-sonorant codas ($t = 1.2791, df = 389, p < 0.2016$). Table 3 lists for both voices the discontinuity detection rates for these two classes in IDFS words.

Class	N Words	Female %	Male %
Sonorant	450	60	62
Non-Sonorant	390	39	36

Table 3: *Female and Male Concatenation Discontinuity Detection Rates and Number of Test Words for IDFS test words with Sonorant and Non-Sonorant Classes of Word Endings*

4. Summary and Discussion

Concatenation discontinuities were significantly more audible overall for the female voice, for which 61.4% of the discontinuities were detected, than for the male voice, for which 54.2% were detected. For both voices, diphthongs had a significantly higher discontinuity detection rate than long or short monophthong vowels. Concatenation discontinuities were significantly more perceptible for both voices when the final halves of the constituents of test words differed (Female: 71.2%; Male: 58.5%) than when the initial halves differed (Female: 50.5%; Male: 50.0%). In the former case (ISFD) for both voices, if either of the two constituent words contained a postvocalic sonorant or either constituent was vowel-final, discontinuity detection rates were significantly higher than if they did not. Even in the latter case when the codas of test word constituents were the same, detection rates for IDFS words with sonorant codas were significantly higher than those with non-sonorant codas for both voices. Pre-vocalic consonants had little effect on detection rates generally. Detection rates for the female voice were consistently higher than the male voice for ISFD test words, but there was no significant overall difference between female and male detection rates for IDFS test words.

Because diphthongs are vowels with two successive targets, they are not marked by a relatively steady-state mid-vowel region as is the case for monophthongs (vowels with a single target). Since all vowels in the study were bisected and concatenated at their mid-points, diphthongs were cut at a spectrally dynamic point during their transition from the first to the second target. Monophthongs, in contrast, were cut within a target region with a relatively stable spectrum. These differences in spectral change at the concatenation point between diphthongs and monophthongs probably account for the concatenation detectability differences observed between the two classes of vowels.

The much larger effect of post-vocalic consonants than of pre-vocalic consonants on the detection of concatenation discontinuities in the vowel indicates that anticipatory (right-to-left) coarticulation has a larger effect than retentive (left-to-right) coarticulation. That is, phones that follow a given phone influence its articulation more or over a longer interval than phones that precede it in time (see [14] for a discussion of this topic). Sonorant consonants have a particularly strong coarticulatory influence on the preceding vowel, which gradually changes during much of its duration. Thus, the mid-point of a vowel preceding a sonorant consonant is relatively less spectrally stable than the mid-point of a vowel preceding a non-sonorant consonant.

The differences in discontinuity detection among vowels and among consonantal contexts consistently suggest that audible discontinuity often results from concatenation in regions of spectral change.

5. References

- [1] A. Conkie and S. Isard, "Optimal coupling of diphones.," in *Progress in Speech Synthesis*, R. Van Santen, R. Sproat, J. Hirschberg, and J. Olive, Eds. 1996, pp. 293–304, Springer Verlag.
- [2] E. Klabbbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *International Conference on Spoken Language Processing ICSLP 98*, pp. 1983–1986, 1998.
- [3] Esther Klabbbers, *Segmental and Prosodic Improvements to Speech Generation*, Ph.D. thesis, IPO, Center for User-System Interaction, 2000.
- [4] E. Klabbbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Proc.*, vol. SAP-09, no. 01, pp. 39–51, Jan 2001.
- [5] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," *International Conference on Spoken Language Processing ICSLP 98*, pp. 2747–2750, 1998.
- [6] A. K. Syrdal, "Perception of consonant place of articulation," in *Speech and Language: Advances in Basic Research and Practice*, N. J. Lass, Ed., vol. 9, pp. 313–349. Academic Press, 1983.
- [7] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.
- [8] J.A. Swets, *Signal detection and recognition by human observers: Contemporary readings*, Peninsula Press, 1988.
- [9] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," in *Proc. Joint Meeting of ASA, EAA, and DEGA*, Berlin, March 1999, ASA, EAA, and DEGA, p. SASCA_4, <http://www.research.att.com/projects/tts/pubs.html>.
- [10] A. S. House, C.E. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A Modified Rhyme Test," Tech. Doc. Rept. ESD-TDR-63-403, U. S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, June 1963.
- [11] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, pp. 158–166, 1965.
- [12] E. J. Kreul, J. C. Nixon, K. D. Kryter, D. W. Bell, J. S. Lang, and E. D. Schubert, "A proposed clinical test of speech discrimination," *J. Speech and Hearing Research*, vol. 11, pp. 536–552, 1968.
- [13] American National Standards Institute, "Method for measuring the intelligibility of speech over communication systems," Revised Standards Report ANSI S3.2-1989 - A revision of ANSI S3.2-1960, American Standards Association, New York, 1989.
- [14] R. D. Kent and F. D. Minifie, "Coarticulation in recent speech production models," *J. Phonetics*, vol. 5, pp. 115–133, 1977.