

Perceptually-based Data-driven Join Costs: Comparing Join Types

Ann K. Syrdal and Alistair D. Conkie

AT&T Labs – Research, Florham Park NJ, USA
syrdal@research.att.com, adc@research.att.com

Abstract

Unit selection synthesis has improved the quality of synthetic speech by making it possible to concatenate speech from a large database to produce intelligible synthesis while preserving much of the naturalness of the original signal. Such synthesis is by no means perfect, however, and this paper describes work to achieve more optimal joins between concatenated units. Results from a psychoacoustic experiment, acoustic parameters and phonetic factors are analyzed and used in statistical training of join costs so that audible discontinuities at concatenation boundaries can be minimized.

1. Introduction

Many modern speech synthesizers use concatenative methods to generate audible speech from text input. The most recent versions of concatenative synthesis, called unit selection [1], use a large inventory of recorded speech in which multiple variants of units are available for concatenation. The first unit selection synthesizer [1] used phones as the minimal units for concatenation, but the introduction of half-phones as minimal units [2] allowed for joins either in the middle of a phone or at a boundary between phones.

Typically in unit selection two cost functions are used in calculating an optimal set of units to form an utterance. The **target cost** is related to evaluating how close (in terms of f_0 , duration and other parameters) a database unit is to a synthesis specification. The **join cost** should be related to the degree of perceived discontinuity. The best overall cost, a weighted sum of target and join costs for various possible speech units, should indicate the sequence of units that give the best quality synthesis. The costs should thus reflect what listeners hear as good or bad synthesis. The relationship between properties of two concatenated units of speech and their perceived discontinuity, the focus of this paper, is important but not well understood.

Most studies relating auditory judgments of concatenation discontinuity to join costs have focused on using perceptual data to evaluate various spectral distance measures – predictive algorithms based on spectral representations of units at the concatenation point [3][4][5][6][7][8]. The goal of these studies was to improve join cost estimates and thus to reduce concatenation artifacts in concatenative synthesis. A few researchers have examined the relationship between phonetic variables and concatenation discontinuities [3][4][5][9][10], and statistically reliable phonetic effects were reported [9][10]. Thus far, all experiments have studied audible discontinuities in mid-vowel joins. Vowels were the initial focus of study because their relatively higher energy was presumed to make concatenation discontinuities more salient.

Recently data-driven methods of determining perceptually-based join costs for concatenative synthesis were introduced [10]. The observed probability of listeners detecting a concate-

nation discontinuity for a given mid-vowel join was used to train join cost models.

The current paper extends the data-driven approach to mid-consonant joins and to consonant-vowel (CV) or vowel-consonant (VC) joins at phone boundaries, and compares the results to those from mid-vowel joins.

2. Perceptual experiment

A rigorous signal detection experiment was conducted on listeners' detection of concatenation discontinuities in words generated by concatenative synthesis from one adult female speaker. Synthesized test words had a single concatenation point, and listeners judged whether or not a test word contained an audible discontinuity.

2.1. Test stimuli

The speech inventory used for test word synthesis consisted of 336 monosyllabic words that constitute the Modified Rhyme Test (MRT) [11], a standard test of speech intelligibility. 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. In several instances, sets contain a word or words that are either vowel-initial or vowel-final. A restricted domain system was built with the MRT inventory using an experimental version of the AT&T unit selection TTS system.

Synthetic test stimuli were synthesized by concatenation of selected portions of the words contained in the acoustic inventory. For mid-vowel joins, each recorded word in the inventory was divided at mid-vowel into two parts. Concatenated test words were synthesized by combining first and second halves of words that contained the same vowel. For the mid-consonant case, a join was either in the middle of the prevocalic consonant or in the middle of the postvocalic consonant. In the phone-boundary case, a join was either at the boundary between the prevocalic consonant and vowel or at the boundary between the vowel and postvocalic consonant. Examples of the various types of joins are listed in Table 1.

A set of 336 control words was also included in the tests. The control words were resynthesized versions of the first and second halves of the same recorded word, and they would be expected to contain no detectable concatenation discontinuities.

The synthesizer was set to use the synchronized overlap add method (SOLA) [12] to concatenate the test words. In this way, concatenation discontinuities due simply to arbitrary abutment of the two halves were avoided, and pitch period continuity was maintained. The original fundamental frequencies of the two constituent word portions was unaltered. Stimuli were sampled at 16 kHz.

Join Type	N	Example
Mid-V: PostV Diff	840	sing + sip = sip
Mid-V: PreV Diff	840	book + look = book
Mid-V: PreV & PostV Diff	873	kith + sing = king
Mid-V: PreV & PostV Same	52	kit + kit = kit
PreV Mid-Consonant	1239	swell + way = sway
PostV Mid-Consonant	962	sent + bend = send
C-V Phone Boundary	300	shook + book = shook
V-C Phone Boundary	240	sing + sill = sill

Table 1: *Types, numbers and examples of concatenated test words*

2.2. Procedure

Rigorous perceptual testing procedures allowed us to use methods of statistical decision theory that have been applied to the general theory of signal detectability [13]. Although the general theory of signal detectability was developed to specify the mathematically optimal detection process, it also has been accepted as a good approximation to a descriptive theory of human detection and recognition behavior, and serves as a guide for the study of human perceptual processes, specifying appropriate experimental methods and statistical treatment of results.

A single interval forced choice Yes/No signal detection paradigm was used in the psychoacoustic experiments. After hearing a test stimulus, a listener reported whether or not (s)he heard a concatenation discontinuity. Each test stimulus was presented once per listener. The entire test battery was divided into a series of subtests; each subtest contained from 45 to 75 test stimuli and normally took from 10 to 15 minutes to complete. Each listener received a different randomization of the stimuli in a subtest. Listening tests were web-based and interactive, and a large majority of listeners used headphones.

Each subtest was composed of both concatenated test words, which had the potential for audible concatenation discontinuities, and control words, which did not. At least one-sixth of the test stimuli in a subtest were control words. Control words provided a means for monitoring of listeners’ false alarm (false positive) errors. A listener’s false alarm rate together with their hit rate (true positive responses) for a subtest determined a d' score, which is a measure of the listener’s sensory capabilities. The parameter d' defines a specific ROC (receiver-operating-characteristic) curve [13].

2.3. Listeners

Forty-five adult volunteer listeners participated in at least one listening subtest. 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. 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; however listener rejection was rarely necessary (a listener was rejected from a subtest less than 4% of the time). There were at least five acceptable listen-

ers for every subtest. The average number of listeners per vowel subtest was 6.4; for phone-boundary subtest, 11.3; and for consonant subtest, 8.0. There were 20,470 total acceptable observations for vowel joins, 19,913 for consonant joins, and 7,373 for phone-boundary joins, totaling 47,756 perceptual judgments for the entire experiment.

3. Acoustic measures

Unlike earlier studies that evaluated spectral distance functions by means of perceptual measures of concatenation discontinuities, in [10] and the current study other acoustic measures were also included as predictor variables. This assumes that a number of factors contribute to the percept of a smooth join, including continuity of f_0 and power as well as spectral similarity across the join.

A set of seven acoustic variables are included in the experiments described here. The choice is based on our previous results from modeling join costs in mid-vowel joins [10]. A compacted set of six vector quantized (VQ) [14] parameters is used. In addition, based on an evaluation of numerous other acoustic parameters, a cross-correlation parameter is included among the acoustic predictor variables.

3.1. Spectral distance

A VQ procedure was applied, with each dataset vector of mel-frequency cepstral coefficients (MFCCs) c_i being labeled as falling into one of 128 categories (variable *cep*). A subset of the data was used to find an initial codebook, and then by successive splitting of all the codebook vectors the codebook was increased by factors of 2 to 128. Each frame in the database was labeled with a VQ codebook value. An appropriately normalized VQ distance table was also calculated.

Two separate codebooks were calculated, one to deal with the standard MFCCs and another codebook to deal with delta MFCCs (d_i ’s), with associated variable *dcep*.

The relevant data frame closest to a unit boundary (in this case the units were half phones) was marked so that when it was time to calculate the concatenation cost between two units, the relevant frame numbers were easily available. Via the frame numbers, the VQ numbers of the abutting frames could be fed into the distance matrix to find two overall “spectral costs”. These costs were between zero and one (one for each codebook).

3.2. Power and f_0

For f_0 and power there were a total of four variables considered. f_0 and power were both extracted from the speech database files at 10ms intervals using a pitch algorithm developed for speech coding (variables f_0 and *pow*). Delta values were derived by differencing adjacent frames, for both f_0 and power (variable names df_0 and *dpow*). The f_0 and df_0 values were rounded to the nearest integer while for the *pow* and *dpow* variables, log values (with a floor) were used rather than raw values. After appropriate scaling, these values were rounded to the nearest integer.

3.3. Cross-correlation

A sequence of samples from a 40ms region round the nominal concatenation point in the speech file is used as data for cross-correlation. Two sequences were obtained, one for each speech segment to be used for the synthesis join. No windowing is

performed on the sequences. The cross-correlation sequence r was calculated with a delay of approximately plus or minus one pitch period from the nominal position. The maximum value for r within the sequence calculated is used as the cross-correlation value for the join cost experiments.

4. Phonetic variables

A classification and regression tree (CART) model [15] using seven phonetic variables were found to account for 30% of the variability in human detection of mid-vowel concatenation discontinuities [10], performing better than models using strictly acoustic variables. The best results were obtained from acoustic-phonetic CARTs, which combined acoustic parameters and phonetic variables. The current paper is primarily focused on acoustic models, although analyses of phonetic effects on concatenation discontinuities and some modeling results incorporating phonetic predictors are presented.

The same seven phonetic variables described in our 2004 study [10] were used in the current experiments. Phonetic variables included a broad phonetic classification of the prevocalic, vocalic, and postvocalic phones in each of the two constituent words plus two binary indicators of whether pre-join and post-join phone pairs were the same or different.

5. Training methods

Linear regression (LR) was the statistical modeling technique used primarily in the current study. Previous work [10] consistently found that LR and optimally pruned CART models performed equivalently in modeling join costs with acoustic predictor variables. In that experiment, each model was trained on the same training data set (80% of the total data set) and tested on the remaining 20% of the total data set. CART models fit the training set very well but were poor predictors of the test set unless they were severely pruned. LR models, on the other hand, performed equivalently well for training and test sets. Consequently, for LR models in the current study the entire data set was used rather than separate sets for training and testing. The output variable of the model was the predicted probability that listeners will hear an audible concatenation discontinuity given a set of predictor variables describing characteristics of the concatenated word.

5.1. Models

The Linear regression (LR) model used in this experiment fits a linear model using the method of least-squares [16]. Linear regression models a numeric response variable, in this case observed detection probability, by a linear combination of numeric predictor variables, in this case, seven acoustic measures. An LR model was trained on each of three data sets representing the three different types of joins: mid-vowel, mid-consonant, and phone-boundary.

A CART model was also used for phonetic and acoustic predictors, as previously described in [10].

6. Join cost evaluation method

Each model's predicted join costs were correlated (using Pearson's product-moment correlation, which yields correlation coefficient r) with observed human detection rates. The squared correlation coefficient r^2 was also calculated; this value represents the proportion of the total variability in the human detection rates that was explained by the predicted join costs.

7. Results

7.1. Phonetic effects on concatenation detection

The previous observation [9] that discontinuities were more likely to be detected in diphthongs than in other broad phonetic classes of vowels was confirmed in the current expanded set of mid-vowel joins. The detection rate of 74.5% for diphthongs was about 15% higher than rates observed in other vowel classes. Table 2 lists listener detection rates for each of the four types of mid-vowel joins tested. If the original postvocalic contexts differed for the two vowels that were joined, the resulting joins were from 16.4% to 28.6% more likely to be detected as discontinuous. This result replicates and extends earlier results [9].

Prevocalic	Postvocalic	% Joins detected
Same	Diff	71.2
Diff	Same	50.5
Diff	Diff	66.9
Same	Same	42.6

Table 2: *Listeners' detection of concatenation discontinuities for four types of mid-vowel joins*

The perceptual salience of mid-consonant joins varied greatly among consonant classes. Table 3 lists the percentage of discontinuities that were detected by listeners for the various broad phonetic categories of consonants studied. Postvocalic consonant joins were detected 49.2% of the time, and prevocalic joins, 46.7%.

Broad phonetic class	% Joins detected
Liquid sonorant	54.7
Glide sonorant	54.7
Aspirated glottal fricative	54.3
Nasal sonorant	47.4
Unvoiced weak fricative	35.3
Voiced weak fricative	23.3
Voiced strong fricative	22.2
Unvoiced strong fricative	13.6

Table 3: *Concatenation detection observed in mid-consonant joins*

In the case of concatenation at phone boundaries, more discontinuities were detected at vowel-consonant boundaries (68.4%) than at consonant-vowel boundary joins (62.0%). Large effects of phonetic context on the detection of discontinuities in phone boundary joins were observed, but they are too numerous to describe in detail here.

The overall detectability of discontinuities in the three classes of joins studied is compared in Table 4.

Join class	% Joins detected
Phone boundary	64.8
Mid-vowel	62.5
Mid-consonant	46.6

Table 4: *Join discontinuity detection for three types of joins*

7.2. Join Cost Prediction

Table 5 lists the r and r^2 values obtained for each of the three linear regression join cost models tested on their own respective type of join and the mid-vowel model tested on the mid-consonant and phone-boundary joins. The last row of the table shows the results obtained for mid-consonant joins from a pruned acoustic-phonetic CART model.

Model	Training set	Test set	r	r^2
LR	Mid-V	Mid-V	0.48	0.23
LR	Mid-V	P-bound	0.30	0.09
LR	P-bound	P-bound	0.52	0.27
LR	Mid-V	Mid-C	0.06	0.00
LR	Mid-C	Mid-C	0.33	0.11
AP-CART	Mid-C	Mid-C	0.56	0.31

Table 5: Test Results for Join Cost Models

8. Discussion and Conclusions

The LR model trained on mid-vowel joins predicted audible discontinuities in mid-vowel concatenated words quite well. Results were comparable to a previously evaluated model [10] that used Euclidean distances on MFCCs, a high-end reference condition that is impractical for real-time unit selection TTS. However, the mid-vowel LR model was a relatively poor predictor of audible discontinuities in either mid-consonant joins or phone-boundary joins.

The phone-boundary LR model performed well in predicting the probability of audible discontinuities in phone-boundary joins. The mid-consonant LR model was relatively less successful, although it was clearly superior to mid-consonant predictions made by the mid-vowel LR model. Acoustic variability among the consonants is probably too wide for a single LR model to be successful. Much better results for mid-consonant joins were achieved by the acoustic-phonetic CART model, where consonants were divided into subsets and different decision criteria used for each.

The relative importance of the seven acoustic predictors varied considerably among the three LR models, as indicated by the t-values and associated probabilities calculated by the linear model. For example, cross-correlation was the best predictor of mid-vowel join discontinuities, but it was third of five significant predictors for phone-boundary joins and was not a significant predictor of mid-consonant joins. Cepstral distance was the best predictor for phone-boundary joins, second of three significant predictors for mid-consonant joins, and fifth of six predictors for mid-vowel joins. Although f_0 was a highly significant predictor for mid-vowel joins, it was not a significant ($p < 0.05$) predictor for either of the other types of joins.

Fundamental research on the perception of concatenation discontinuities and use of perceptual data to train join cost functions provide a disciplined empirical approach to improving the quality of concatenative synthesis.

9. References

- [1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 373–376, 1996.
- [2] A. Conkie, "A robust unit selection system for speech synthesis," *137th meeting of the Acoustical Society of America*, p. 978, 1999.
- [3] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *International Conference on Spoken Language Processing ICSLP 98*, pp. 1983–1986, 1998.
- [4] E. Klabbers, *Segmental and Prosodic Improvements to Speech Generation*, Ph.D. thesis, IPO, Center for User-System Interaction, 2000.
- [5] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Proc.*, vol. SAP-09, no. 01, pp. 39–51, Jan 2001.
- [6] 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.
- [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] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers," *Fourth ESCA Speech Synthesis Workshop*, 2001.
- [9] A. K. Syrdal, "Prosodic effects on listener detection of vowel concatenation," *Proc. EUROSPEECH*, 2001.
- [10] A. K. Syrdal and A. D. Conkie, "Data-driven perceptually-based join costs," *Fifth ISCA ITRW on Speech Synthesis (SSW5)*, pp. 49–54, 2004.
- [11] 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.
- [12] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1985, pp. 493–496.
- [13] J.A. Swets, *Signal detection and recognition by human observers: Contemporary readings*, Peninsula Press, 1988.
- [14] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1587, Nov 1995.
- [15] L. Breiman, J. H. Friedman, R. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.
- [16] S. R. Searle, *Linear Models*, J. Wiley & Sons, New York, 1971.