

DATA-DRIVEN PERCEPTUALLY BASED JOIN COSTS

Ann K. Syrdal and Alistair D. Conkie

AT&T Labs – Research, Florham Park, NJ, USA

ABSTRACT

Concatenative speech synthesis systems attempt to minimize audible discontinuities between two successive concatenated units. In unit selection concatenative synthesis, a join cost is calculated that is intended to predict the extent of audible discontinuity introduced by the concatenation of two specific units. A study was conducted that used human perceptual data on the detectability of mid-vowel concatenation discontinuities to train and to test several models for predicting perceptually-based join costs. Both linear regression (LR) and classification and regression tree (CART) models were used. Each was trained on several different sets of predictor variables. All LR and some CART models used strictly acoustic predictor variables, some CART models used acoustic plus phonetic categorical variables, and one CART model used strictly phonetic predictors. Results from tests of LR and CART models showed that, when trained with the same acoustic predictor variables, the two models achieved very similar results in predicting human detection rates. Euclidean cepstral distances were superior to VQ cepstral distances as predictor variables. Categorical phonetic predictor variables in CART models greatly improved the accuracy of prediction of concatenation discontinuities.

1. INTRODUCTION

Many popular speech synthesizers use concatenative methods to generate audible speech from text input. Concatenative synthesizers have a database of recorded speech from which suitable fragments are selected and concatenated (joined together) to produce a desired utterance. In diphone synthesis [1], an acoustic inventory is composed of recorded speech fragments, each representing the transition from the quasi-steady-state center of one phone to the center of the following phone; this unit constitutes a diphone. Diphone inventories are often limited to a single speech fragment for each diphone that can occur within or between words in a given language. In a more recent method of concatenative synthesis called unit selection [2], a much larger inventory of recorded speech is used in which multiple variants of units are available for concatenation.

Unit selection synthesis has the potential for higher quality and more natural sounding synthetic speech, but it also requires an algorithm to select at run time the most appropriate units available to construct the desired utterance. Typically two cost functions are used in calculating an optimal set of units to form an utterance. One cost (target cost) is related to evaluating how close a database unit is to a synthesis specification. Are the f_0 , duration and other parameters of a database unit a good match for the requested unit? A low cost represents a good match. The second cost (concatenation cost or join cost) is intended to indicate the quality of a concatenation between two adjacent units. The join cost should be low if a concatenation is likely to be smooth, and high if it is likely to be perceived as an annoying discontinuity. The minimum overall cost, a weighted sum of target and concatenation costs for various possible synthesis units, should indicate the sequence of units that give the best quality synthesis. For this to happen it is important that the costs reflect, as far as possible, what listeners hear as good or bad synthesis. However, the relationship between properties of two concatenated units of speech and their perceived discontinuity, the focus of this paper, is one important area that is 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]. Many of the spectral distance measures evaluated and compared in these experiments were motivated by those used in speech recognition methods. For synthesis purposes they can be used to calculate join costs. The goal of these studies was to improve join cost estimates and thus to reduce concatenation artifacts in concatenative synthesis. Few studies have examined the relationship between phonetic variables and concatenation discontinuities [3][4][5][9], although statistically reliable phonetic effects were reported [9]. Thus far, all experiments have studied audible discontinuities in vowel joins. Vowels are important as an initial focus of study because their relatively higher energy makes concatenation discontinuities more salient.

This paper introduces new data-driven methods of determining perceptually-based join costs for concatenative

synthesis. The observed probability of listeners detecting a concatenation discontinuity (listeners' detection rate) for a given join was used to train join cost models. Several combinations of training models and sets of predictor variables were studied and compared. A database of 2605 concatenated test words (from one female speaker), each with a single mid-vowel join was used as the basis of the experiments. The observed listener detection rate of each concatenated word join plus several acoustic measures and phonetic features of the two constituent speech fragments were used in training and testing. Eighty percent (2084 concatenated words) of the database was used for training, and the remaining 20% (521 concatenated words) for testing.

2. PERCEPTUAL EXPERIMENT

A rigorous psychoacoustic signal detection [10] experiment was conducted on listeners' detection of concatenation discontinuities in words generated by concatenative synthesis using speech data from one adult female speaker. Synthesized test words had a single concatenation point, located in the middle of the vowel, 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][12][13], a standard test of speech intelligibility [14]. 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 (such as in "oil, foil, coil, boil, toil, soil") or vowel-final (such as in "ray, raze, rate, race, rake, rave"). A restricted domain system for each voice was built with the MRT inventory using an experimental version of the AT&T unit selection TTS system. The MRT inventory of recorded isolated words was selected for the experiment because it provided many opportunities for synthesizing test words with only mid-vowel joins. Restricting test words to one concatenation point was necessary so that an audible discontinuity could be attributed only to the single join in the synthesized word.

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 at mid-vowel 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).

Four types of concatenated test words were synthesized by combining first and second halves of words that contained the same vowel. The type classification depended upon differences between the two prevocalic contexts and between the two postvocalic contexts of the two words from which the concatenated halves were taken. Table 1 lists the four concatenated test word types, the number and an example of each type.

Prevocalic	Postvocalic	N	Example
Same	Diff	840	sing + sip = sip
Diff	Same	840	book + look = book
Diff	Diff	873	kith + sing = king
Same	Same	52	kit + kit = kit

Table 1. Four Types of Concatenated Test Words

A set of 336 control words was also included in the test. 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) [15] to concatenate the first and second halves of words at approximately the mid-point of the vowel. 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 halves was unaltered. Stimuli were sampled at 16 kHz.

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 [10]. 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. Adherence to this theoretically and methodologically well-grounded foundation is a major strength of the present study, and is unique in this respect among the perceptual studies of concatenation. Some of the other studies have reported difficulties in obtaining (or in determining whether or not they have obtained) reliable perceptual data, and some have rejected a high proportion of listeners tested.

A single interval forced choice Yes/No signal detection paradigm was used in the psychoacoustic experiments. Af-

ter 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. Typically, a listener would participate in no more than one subtest a day. Written instructions to listeners and several examples of a stimulus in 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.

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. The inclusion of control words in each subtest provided a means for the careful monitoring of listeners' false alarm (false positive) errors. A listener's false alarm rate together with their hit rate (true positive responses) determined their d' score for that subtest. The variable d' is a measure of the listener's sensory capabilities (as distinct from the listener's performance), and the parameter d' defines a specific ROC (receiver-operating-characteristic) curve [10].

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 listeners for every stimulus word in the test set. The average number of listeners per subtest was 6.4. There were 20,470 total acceptable observations in the experiment.

3. ACOUSTIC MEASURES

Unlike the previously cited studies that evaluated spectral distance functions by means of perceptual measures of concatenation discontinuities, in the current study other acoustic measures are also included as predictor variables. It seems reasonable to assume 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.

In the experiments described here a number of acoustic variables are considered. They fall into two overlapping classes. The first class contains a full set of cepstral parameters, and also includes f_0 and power values. Use of a full set of parameters is primarily for experimental purposes, maximizing the number of features available for analysis. These parameters are not intended for, nor are they useful for general-purpose run-time synthesis, for two reasons. First, storage of a large number of such variables would significantly increase memory usage. Secondly calculations involving a large number of parameters would be significantly slower than the typical concatenation cost approximation used in practice. Given that the overall concatenation cost calculation time is (very approximately) proportional to the square of the beam width of the best path calculation, using the unprocessed parameters would be dramatically slower than what can be achieved with processed parameters. Consequently, for the purposes of achieving synthesis speed a compacted set of parameters is often used (6 in total), Vector Quantized (VQ) [16] in order to enhance the speed of calculation.

3.1. Spectral Distance

In a previous study [7] conducted with a subset of the database of the current study, the Euclidean distance on Mel-Frequency Cepstral Coefficients (MFCCs) was found to be one of the two best spectral distance measures among 13 evaluated as predictors of audible concatenation discontinuities. Consequently, Euclidean distances on MFCCs and their delta coefficients were included, in combination with other predictor variables, as a reference in join cost estimates.

MFCCs were derived from raw speech signals at 10ms intervals. The number of coefficients used as predictor variables in these experiments was 24: 12 cepstral coefficients and 12 delta cepstral coefficients. The cepstral coefficients were derived using standard methods commonly used in speech recognition [17]. The entire speech database was processed in this way. For the purely experimental aspects

of the study, where the full parameter set was used, the MFCCs were used directly. So the coefficients were, c_i , where $1 \leq i \leq 12$ and for the deltas d_i , where $1 \leq i \leq 12$.

For the compacted version of the experimental variables, a VQ procedure was applied, with each dataset vector of coefficients 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*.

For both the full cepstral set and the vector quantized set 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 in the case of the vector quantized values (one for each codebook), or a set of absolute differences between individual coefficients for the unquantized case.

3.2. f_0 and Power

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 standard 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. Predictor Variable Sets

Three sets of acoustic variables were used as predictor variables for training models in this study:

Set A30. 30 predictors: *cep*, *dcep*, f_0 , df_0 , *pow*, *dpow*, MFCCs c_1 - c_{12} , and their delta coefficients d_1 - d_{12}

Set A6C. 6 predictors: *cep*, *dcep*, f_0 , df_0 , *pow*, *dpow*

Set A6E. 6 predictors: Euclidean distances on MFCCs c_1 - c_{12} and on d_1 - d_{12} , f_0 , df_0 , *pow*, *dpow*.

4. PHONETIC VARIABLES

Since statistically reliable phonetic effects on concatenation discontinuities were previously observed [9], seven phonetic

variables (termed **Set P7**) were also included in some statistical models as predictors of join costs. A broad phonetic classification of the vowel being joined, of its preceding phone, and of its subsequent phone in each of the two constituent words provided five categorical predictor variables.

Consonants were classified into ten broad phonetic manner categories: Glide Sonorants, Liquid Sonorants, Nasal Sonorants, Aspirated Glottal Fricatives, and Voiced and Voiceless cognates of Plosives, Strong Fricatives, and Weak Fricatives. Vowels were classified into five broad phonetic categories: Diphthongs, Long Front Vowels, Long Back Vowels, Short Front Vowels, and Short Back Vowels.

In addition, two binary categorical predictors were used to indicate whether the prevocalic phones of the two constituent words were the same or different, and whether the postvocalic phones of the two constituent words were the same or different.

5. TRAINING METHODS

Two different basic statistical models were used in this experiment: linear regression (LR) and classification and regression trees (CART). Each model was trained on the same training data set (80% of the total data set), using several different sets of predictor variables, and each was tested on the same test data set (the remaining 20% of the total data set). For each model, the output variable was the predicted concatenation detection rate for a given concatenated word (i.e. the probability that listeners will hear an audible concatenation discontinuity given a set of predictor variables describing characteristics of the concatenated word).

5.1. Linear Regression (LR)

The LR model used in this experiment fits a linear model using the method of least-squares [18]. Linear regression models a numeric response variable, in this case observed detection rate, by a linear combination of numeric predictor variables. Each of the variables was observed on the same 2084 words that made up the training data set.

An LR model was trained on each of the three different sets of acoustic predictor variables described in section 3.3. Their code names reflect the predictor variable set used for their training: LR A30, trained on set A30; LR A6C, trained on set A6C; and LR A6E, trained on set A6E.

5.2. Classification and Regression Trees (CART)

Tree-based models are an alternative to linear and additive models for regression and to linear and additive logistic models for classification [19]. Tree-based models are fitted by binary recursive partitioning that successively splits a data set into increasingly homogeneous subsets. An advantage

of the CART approach is that predictor variables may be either numeric or categorical. Since CARTs recursively partition a data set until it is infeasible to continue, they tend to over-fit to the training data, so that full CARTs are less robust and generalizable than simplified CARTs pruned to include only the higher decision nodes in the tree. For this reason, only pruned versions of each CART were included in the experiment. The pruned CARTs described below were those whose predicted joint costs correlated most highly with observed detection rates of the test set.

Three CART models were trained on the same three sets of acoustic predictor variables used for the LR models, and pruned. Again, their code names reflect the predictor set used to train them: (1) CART A30, trained on set A30, for which 7 of 30 predictors remained after pruning; (2) CART A6C, trained on set A6C, for which 5 of 6 predictors remained after pruning; (3) CART A6E, trained on set A6E, of which 5 of 6 predictors remained after pruning.

Two acoustic-phonetic CART models were trained on six numeric acoustic plus seven categorical phonetic predictors, and pruned: (1) CART A6C_P7 was trained on acoustic set A6C and 7 phonetic predictors, of which 2 of 6 acoustic predictors and 6 of 7 phonetic predictors remained after pruning; and (2) CART A6E_P7, trained on acoustic set A6E and 7 phonetic predictors, of which 3 of 6 acoustic predictors and 4 of 7 phonetic predictors remained after pruning.

One additional CART model, CART P7, was trained only on the seven categorical phonetic predictor variables, for which 3 of 7 phonetic predictors remained after pruning.

6. JOIN COST EVALUATION METHODS

Each model’s predicted joint costs were correlated (using Pearson’s product-moment correlation, which yields correlation coefficient r) with the observed human detection rates for both the training data set and the testing data set. 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 joint costs.

7. RESULTS

Table 2 lists the r and r^2 values obtained for each join cost model for both training and testing data sets. Discussion of results will focus on performance of the various models for the test set.

7.1. Acoustic Models

LR A30 and LR A6E had very similar results and were clearly the best of the three LR models evaluated. Of the LR

Join Cost Models	Training Set		Testing Set	
	r	r^2	r	r^2
LR A30	0.50	0.25	0.49	0.24
LR A6C	0.44	0.19	0.42	0.18
LR A6E	0.48	0.23	0.50	0.25
CART A30	0.51	0.26	0.47	0.22
CART A6C	0.49	0.24	0.43	0.19
CART A6E	0.52	0.27	0.51	0.26
CART A6C_P7	0.66	0.44	0.56	0.31
CART A6E_P7	0.67	0.45	0.57	0.33
CART P7	0.59	0.32	0.55	0.30

Table 2. Training and Test Results for Join Cost Models

models with 6 predictor variables, LR A6E was clearly superior to LR A6C. Both of these results indicate the success of the Euclidean distance measure as an effective predictor variable.

CART A6E was the best performing of the acoustic CART models evaluated, with results similar to the two best LR models, LR A30 and LR A6E. As with the LR models, the CART model using six predictor variables that included two Euclidean distances, CART A6E, was clearly superior to its six predictor counterpart with VQ *cep* distances instead, CART A6C.

In general, LR and acoustic CART models that were trained with the same set of acoustic predictor variables yielded comparable results. Both the LR and the pruned CART models that used the predictor variable set E6 explained 7% more of the total variability in human detection rate than the models that used the C6 set of predictor variables.

7.2. Acoustic-Phonetic Models

CART A6E_P7 had only slightly higher test results than CART A6C_P7. Thus, for the acoustic-phonetic models, the difference between the A6E and A6C predictor variable sets was in the same direction but not nearly so large as for the acoustic models using either LR or CART.

The addition of seven phonetic categories to either of the two sets of six acoustic predictor variables greatly improved the CART models’ test performance. The biggest gains resulted from adding phonetic variables to the A6C predictor variable set: CART A6C_P7 explained 12% more of the variability in human detection rates than CART A6C, and 11% more than explained by LR A6C. Similarly but on a smaller scale, CART A6E_P7 explained 7% more detection rate variability than CART A6E, and 8% more than LR A6E. The two acoustic-phonetic CART models were the two best join cost predictors evaluated.

7.3. Phonetic Model

Trained with only the seven categorical phonetic predictor variables, CART P7 performed nearly as well as the acoustic-phonetic CARTs, explaining 30% of the variability in human detection rates. Its performance was higher than that of any of the acoustic models tested.

8. CONCLUSIONS AND DISCUSSION

Consistent patterns of results from these experiments lead to a number of conclusions: (1) Euclidean distance measures were superior to VQ cepstral distances (with codebook size 128) as predictors of audible concatenation discontinuities; (2) LR and pruned CART models were essentially equivalent in predicting human detection rates from acoustic predictor variables; (3) the use of categorical phonetic predictor variables in CART models, either alone or in addition to acoustic variables, greatly improved the prediction of human concatenation detection.

The success of phonetic variables in this study seems to indicate a wider contextual influence on concatenation discontinuities than that encompassed by the acoustic variables studied.

9. REFERENCES

- [1] G. Peterson, W. Wang, and E. Sivertsen, "Segmentation Techniques in Speech Synthesis," *J. Acoust. Soc. Am.*, vol. 30, pp. 739–742, 1958.
- [2] 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.
- [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] Esther 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] J.A. Swets, *Signal detection and recognition by human observers: Contemporary readings*, Peninsula Press, 1988.
- [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] 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.
- [13] 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.
- [14] 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.
- [15] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 493–496, 1985.
- [16] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, pp. 1551 – 1587, Nov 1995.
- [17] L. R. Rabiner and B-H. Juang, *Fundamentals of speech recognition*, PTR Prentice-Hall, 1993.
- [18] S. R. Searle, *Linear Models*, J. Wiley & Sons, New York, 1971.
- [19] L. Breiman, J. H. Friedman, R. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.