The Contribution of Various Sources of Spectral Mismatch to Audible Discontinuities in a Diphone Database

Esther Klabbers, Jan P. H. van Santen, *Member, IEEE*, and Alexander Kain

***Abstract*—One of the major problems in concatenative synthesis is the occurrence of audible discontinuities between two successive concatenative units. Several studies have attempted to discover objective distance measures that predict the audibility of these discontinuities. In this paper, we investigate mid-vowel joins for three vowels with a range of post-vocalic consonant contexts typical for diphone databases. A first perceptual experiment uses a pairwise comparison procedure to find two subsets of unit combinations: Those *with* versus *without* audible discontinuities. A second perceptual experiment uses these two subsets in a procedure where formant resynthesis is used to manipulate three sources of discontinuity separately: formant frequencies, formant bandwidths, and overall energy. Results show mismatch in formant frequencies provides the largest contribution to audible discontinuity, followed by mismatch in overall energy.**

***Index Terms*—Audible discontinuities, diphones, spectral distance measures, speech synthesis.**

I. Introduction

UNIT selection synthesis is currently the most popular synthesis method [1], [2]. This process consists of searching a large database of speech units for the optimal sequence of units, employing a range of cost functions. Diphone synthesis is a simplified form of unit selection where there is generally only one instance of each diphone unit in the database and the concatenated units are usually selected from different utterances. One of the most complicated problems in concatenative synthesis is the occurrence of audible discontinuities at concatenation boundaries. Numerous studies have been dedicated to finding objective distance measures that will predict this phenomenon [3]–[10]. It has been shown that certain distance measures are better at predicting audible discontinuities than others, but the correlations between distance measures and perceptual quality are generally weak. Donovan reports correlations between 0.05 and 0.4 with one outlier of 0.6 [3]. Wouters *et al.* report correlations between 0.28 and 0.66 [10]. Furthermore, studies are difficult to compare because there is neither a standard corpus for this purpose nor a standard concatenation operation. Nevertheless, we can draw several important conclusions from these previous studies:

- Listeners should rate one cut point at a time: When comparing two sentences generated by a unit selection synthesizer with different join costs, the number of cut points in each sentence can vary, and the effect of one bad cut point in one sentence can heavily influence the decision, especially when that cut point occurs towards the end of the sentence.
- Provide a reference stimulus: A pairwise experiment in which exactly one word has a possible mismatch makes it easier for listeners to reliably detect discontinuities. Without a reference, listeners may use preceding stimuli as their reference, thus adding uncontrolled variability to the results.
- Separate spectral sources of mismatch from prosodic sources: Some systems use prosodic modification after the unit selection process and others do not. Focusing on only one source of mismatch will simplify the task.

The current study attempts to determine the relative contributions of several potential sources of mismatch to audible discontinuity. It is carried out in the context of recordings made for a diphone database. This means that the recordings are highly structured and contain mostly systematic variation due to phonetic context, and not prosodic context. We focus on the mismatch in formant frequencies, formant bandwidths, and overall energy. Two experiments are performed. In Section II, we present the first perceptual experiment that uses a pairwise comparison procedure to find two subsets of unit combinations: those *with* versus those *without* audible discontinuities. The results from this experiment are correlated with different objective distance measures to gain insight into the role of each of the sources of mismatch. In Section III, we present the second experiment that uses these two subsets in a procedure where formant resynthesis is used to manipulate three sources of discontinuity separately: formant frequencies, formant bandwidths, and overall energy. Section IV presents a conclusion and Section V a discussion.

II. Detection Experiment

A psychoacoustic experiment was conducted on listeners' detection of concatenation discontinuities in a large number of modified words. These words were generated by concatenative synthesis using nonsense words that were used to construct a diphone database, as recorded by a female speaker. The reason for choosing these recordings is that the speech is pronounced in a highly structured fashion, leaving mostly phonetic variation caused by coarticulation (which appears systematically) and speaker variability (which appears randomly). Moreover, because the stimuli are short and contain only one cut point,

Manuscript received January 5, 2006; revised July 14, 2006. This work was supported by the National Science Foundation under Grant 0313383: "Objective Methods for Predicting and Optimizing Synthetic Speech Quality." The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

The authors are with the Center for Spoken Language Understanding, OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR 97206 USA (e-mail: klabbers@cslu.ogi.edu; vansanten@cslu.ogi.edu; kain@cslu.ogi.edu).

Digital Object Identifier 10.1109/TASL.2006.885250

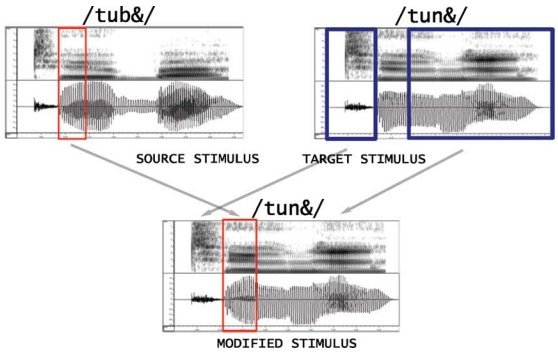

Fig. 1. Example of stimulus creation. The first half of the vowel in /tub&/ is inserted into the word /tun&/.

subjects will be better able to detect a discontinuity if there is one present.

A. Test Stimuli

The speech data consists of words of the form /tVC&/ (e.g., *tuna*, /tun&/), taken from a diphone inventory of a native American English female professional speaker. The speech signals were recorded and stored in 16-kHz 16-bit PCM format. To reduce the number of stimuli, the selection of the vowel V was restricted to be one of three vowels /i:/, /@/, and /u/.[1] These three vowels occur at the extremes of the vowel triangle. Moreover, it only included cases where the prevocalic consonant was constant (/t/) and the postvocalic consonant C was one of 24 consonants /p, t, k, b, d, g, m, n, N, f, T, s, S, h, v, D, z, Z, tS, dZ, l, 9r, j, w/. Previous research by Syrdal has shown that post-vocalic consonants influenced detection of concatenation discontinuities significantly more than pre-vocalic consonants [6]. The spectral distances are generally larger because of an anticipatory coarticulation effect.

Because the recordings came from a diphone inventory, the variation in pitch and duration was limited. In order to eliminate the possibility of pitch discontinuities completely and to make the stimuli comparable in a pairwise experiment, the vowel duration was normalized to its average duration for that vowel as found in the diphone database (145 ms for /i:/, 185 ms for /@/ and 125 ms for /u/) and the pitch was monotonized to an average value of 220 Hz using STRAIGHT [12]. This manipulation did not cause any audible degradation of the signal, because the modification factor was close to 1.0.

The stimulus creation process is illustrated in Fig. 1. The cut point was determined by the pitch mark nearest the mid-point of the vowel region. The first half of the vowel in the source stimulus was then transplanted into the target stimulus creating the modified stimulus, i.e., the first half of the /u/ in /tun&/ was replaced by the first half of the /u/ in /tub&/. We decided to keep the /t/ the same in both the modified and reference stimulus to avoid differences in explosion and frication in the /t/ which could affect the discontinuity ratings. Because there is a clear and abrupt transition between the /t/ and the vowel onset, the concatenation at this boundary was a straightforward abutment. This process did not create any additional discontinuities at the /t/-V boundary, because all vowels were originally preceded by a /t/ and because the abutment took place at a zero-crossing right after the /t/. The process was applied to every single stimulus so as not to introduce a confounding variable. We used a simple overlap-add routine at the center of the vowel, using the center pitch mark and the two surrounding ones to do a linear cross-fade. A reference stimulus was created in which only the pitch and duration were altered to match the modified stimulus. The modified stimulus and the reference stimulus could then be compared side-by-side.

[1]Phoneme names are expressed in Worldbet, an ASCII version of IPA [11].

TABLE I
FIVE -POINT PERCEPTUAL SCALE FOR CMOS

-2. A sounds better
-1. A sounds slightly better
0. About the same
1. B sounds slightly better
2. B sounds better

There were 24 words per vowel that each had a different post-vocalic consonant. Each word could serve as the source and as the target stimulus which means that there are $24 * 24 = 576$ possible combinations. However, 24 of these are cases where the source and target stimulus are the same, creating the reference stimulus, leaving 552 modified stimuli per vowel. For three vowels that produces a total of 1656 modified stimuli. Because each listener was to be presented with all stimuli, a technique was sought to further reduce the number of stimuli, in order to keep the length of the experiment suitably short to avoid subject fatigue. If one were to create a table with all the consonant contexts for the source stimuli in the rows and all the consonant contexts for the target stimuli in the columns, only those cases were selected where the row index was greater than the column index. Thus, only the top right triangle of the table was filled. The resulting set contained 276 stimuli per vowel, or 828 total.

B. Procedure

The stimuli were presented in a Comparative Mean Opinion Score (CMOS) test based on the guidelines in ITU P.800 [13]. Six expert listeners listened to pairs of utterances and were asked to rate the quality of stimulus "A" relative to stimulus "B." One of them was the modified stimulus and the other was the reference stimulus. This provided listeners with an anchor point, which, as remarked in the introduction, was expected to make the ratings easier and more reliable. A five-point perceptual scale was used as shown in Table I. The order of A and B was randomized. Half the stimuli were presented in "AB" order and half in "BA" order. The stimulus list was randomized and divided into three blocks. Six lists were created with block orders {1, 2, 3}, {2, 3, 1}, {3, 1, 2} and stimuli that were presented in "AB" order were presented in "BA" order in the second list with the same block order.

The experiment was supplemented with two stimuli for validity testing, one where the modified stimulus was the same as the reference stimulus, and one where the modified stimulus had the two halves coming from different vowels. In addition, one stimulus was repeated ten times throughout the experiment for reliability testing. Each block started with six examples which were excluded from analysis, one good and one bad example for each vowel.

The experiment was administered to the listeners using WWStim [14], a CGI-based script that automatically presents auditory stimuli to the listener accompanied by the five-point scale mentioned above, implemented as radio buttons. The test was performed on one computer in the Center for Spoken Language Understanding (CSLU) Perception Laboratory, which is equipped with an M-Audio Duo USB Audio interface and a

high-quality AKG head set. The listeners were allowed to listen to the stimulus pairs multiple times before making a decision, but once a decision was made, the listeners could not go back. They could only listen to pairs of stimuli, not to each stimulus individually. Because of the large number of stimuli each listener had to rate, the experiment was split up into three sessions of 280 stimuli each, which were completed on different days. Each session took between 30 and 40 min to complete.

C. Analysis

The scores were first transformed such that they reflected the "AB" order, where "A" is the modified stimulus and "B" is the reference stimulus. Thus, a higher score corresponded to a preference for the reference stimulus, due to a perceived discontinuity in the modified stimulus. We computed the final score for each stimulus by calculating the average score over all subjects. Originally we computed a principal component analysis (PCA) score which assigned larger weights to listeners who were in general agreement with each other and lower weights to outliers. However, the correlation between the average subject score and the PCA score was so high (0.99) that the average score proved sufficient for our purposes. We will use this average score to select the 20 best and 60 worst cases of the experiment. We selected more bad cases than good cases because we expected different kinds of discrepancies between the vowels to create discontinuities. We need more discontinuous stimuli to analyze these discrepancies properly. These will form the basis for a second perceptual experiment in which formant frequencies, formant bandwidths and overall energy are modified one at a time. The average scores were correlated with different objective distance measures to determine how well these measures can predict the audible discontinuities.

1) Formant Frequencies: Formant frequencies can be measured on different scales. The most straightforward representation is the linear scale in hertz as measured by the ESPS utility *formant* which is available through the Snack toolkit [15]. The linear scale can be transformed to a log scale to correspond closer to human perception. The Bark scale [16], [17] is a more sophisticated scale that consists of a number of critical bands. The bandwidth of each band is equal up to 700 Hz and approximately 1/3 octave above that. The ERB scale [18] provides another way to represent the frequencies. ERB stands for "equivalent rectangular bandwidth" It is closely related to the log and the Bark scale. Table II represents the results of performing a linear regression on the average subject scores using the different formant frequency distance measures as predictors. R^2 represents the variance explained by the linear regression model. The F-value signifies whether the model as a whole has statistically significant prediction capability. The Sig(nificance) shows that all models are significant at the 0.001 level. The linear representation does not correlate as well with the average subject score as the other measures. The log scale representation is better, but not as good as the ERB and Bark scales. The ERB and Bark scales yield similar results. The correlation between the two is 0.987. Since the Bark scale is a well-established scale, it is used in the analyses that follow.

Weighted Euclidean formant frequencies were created using linear regression coefficients obtained from regressing the average subject score with the distances for each of the three formants separately. The R^2 for the linear scale improved to 0.17, which was still the lowest score. The impact on the other scales was negligible. The weights were different for each measure, as they were determined by the linear regression coefficients obtained for each measure. The R^2 for the weighted Bark scale improved to 0.18, which given the fact that it added three extra parameters to the analysis, did not warrant the inclusion of the weights. The weights from the regression analysis of the linear scale decreased with each higher formant, confirming the validity of the negative acceleration present in all standard transformations, whether log, Bark, or ERB. We also calculated the correlation between the average subject score and each formant frequency separately, which also confirmed these findings.

TABLE II
LINEAR REGRESSION RESULTS FOR DIFFERENT FORMANT FREQUENCY DISTANCES. R^2 REPRESENTS THE VARIANCE EXPLAINED BY THE LINEAR REGRESSION MODEL. SIG. REPRESENTS THE SIGNIFICANCE WHERE $p < 0.001$ IS $***$, $p < 0.01$ IS $**$, AND $p < 0.05$ IS $*$

	R^2	F-value	Sig.
$D_{FF}(Hz)$	0.074	66.0	$***$
$D_{FF}(logHz)$	0.16	156.5	$***$
$D_{FF}(ERB)$	0.17	171.7	$***$
$D_{FF}(Bark)$	0.17	169.3	$***$

To summarize, the analysis includes the Euclidean formant frequency distance in Bark as presented in (1). The index k refers to the formant, and the subscripts l and r refer to the left and right side of the join. The formants were measured using the ESPS utility *formant*. The formants were then transformed to the Bark domain using Traunmüller's refinement [17], which reduces over- and underestimation at the extremes. From correlating each of the formants separately with the experiment results, we concluded that the fourth formant did not contribute significantly. The distances were measured at the cut point in the source and target stimulus

$$D_{\mathrm{FF}}(l,r) = \sqrt{\sum_{k=1}^{3}(\mathrm{FF}_{k,l} - \mathrm{FF}_{k,r})^2}. \tag{1}$$

2) Formant Bandwidths: The contribution of formant bandwidth mismatch is quantified by the Euclidean distance between formant bandwidths. The bandwidths are also obtained via the ESPS *formant* function and transformed to the Bark domain using (2). This is necessary because the Bark scale is nonlinear and transforming the bandwidth values into the Bark scale directly would result in overestimation of the bandwidth values with larger errors for the higher bandwidths. Formant bandwidths are not likely to be reliably estimated by a formant tracker, especially if the tracking is based on LPC poles and the bandwidth of the nearest pole is taken as the formant bandwidth. We performed a visual inspection of the calculated bandwidths and removed outliers to create smooth bandwidth tracks

$$\mathrm{F2B}\left(\mathrm{formant}_i + \left(\frac{\mathrm{bandwidth}_i}{2}\right)\right) - \mathrm{F2B}\left(\mathrm{formant}_i - \left(\frac{\mathrm{bandwidth}_i}{2}\right)\right) \tag{2}$$

$$D_{\mathrm{FB}}(l,r) = \sqrt{\sum_{k=1}^{3}(\mathrm{FB}_{k,l} - \mathrm{FB}_{k,r})^2}. \tag{3}$$

3) Overall Energy: The contribution of overall energy mismatch is quantified by the Euclidean energy distance [(4)]. The energy is computed by taking the mean energy of a low-passed

speech signal over a triangular window that spans between the two pitch marks immediately to the left and right of the join

$$D_{\mathrm{E}}(l,r) = \sqrt{(\mathrm{E}_l - \mathrm{E}_r)^2}. \quad (4)$$

4) Symmetrical Kullback–Leibler Distance (SKL): In addition to these measures, two measures were included that have frequently been used in other studies, the symmetrical Kullback–Leibler measure and the Euclidean distance on Mel-cepstral coefficients. The symmetrical Kullback–Leibler measure [(5)] was the best measure in a perceptual experiment reported in Klabbers and Veldhuis [4]. It has been adopted in many studies since, with fairly good results. It computes the distance between two power-normalized LPC spectral envelopes. It has the important property that it emphasizes differences in spectral regions with high energy more than differences in spectral regions with low energy. Spectral peaks are thus more important than valleys between the peaks, and low frequencies are more important than high frequencies. The power-normalized spectral envelopes $P(\omega)$ and $Q(\omega)$ are computed over a 40-ms window around the pitch mark nearest the center of the vowel

$$D_{\mathrm{SKL}}(P,Q) = \int (P(\omega) - Q(\omega)) \log\left(\frac{P(\omega)}{Q(\omega)}\right) \mathrm{d}\omega. \quad (5)$$

5) Mel-Frequency Ceptral Coefficients (MFCCs): The Euclidean MFCC distance, or D_{MFCC}, is computed as in (6). The order p is set to 13. The first MFCC coefficient is not included in the analysis as it represents the overall energy, which is already taken into account by the overall energy distance

$$D_{\mathrm{MFCC}} = \sum_{k=1}^{p} (c_{k,l} - c_{k,r})^2. \quad (6)$$

TABLE III
CORRELATIONS BETWEEN SUBJECTS AND THE AVERAGE SCORE

	subj1	subj2	subj3	subj4	subj5	subj6
ave	0.52	0.66	0.47	0.61	0.49	0.66
subj1		0.20	0.09	0.19	0.10	0.21
subj2			0.19	0.24	0.19	0.33
subj3				0.20	0.06	0.27
subj4					0.18	0.26
subj5						0.16

TABLE IV
CORRELATIONS BETWEEN OBJECTIVE DISTANCE MEASURES

	D_{SKL}	D_{MFCC}	D_{FB}	D_{E}
D_{FF}	0.65	0.64	0.16	0.38
D_{SKL}		0.69	0.34	0.29
D_{MFCC}			0.19	0.49
D_{FB}				-0.05

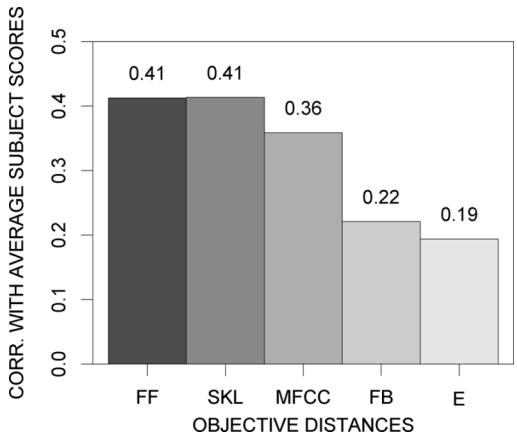

Fig. 2. Correlation between objective distance measures and average subject score.

D. Results

The experiment contained two stimulus pairs that served to test the validity of the experiment. In one pair, the test and reference stimulus were the same; in the other pair, the modified stimulus contained vowel halves from two different vowels, /i:/ and /u/. This presented such a mismatched case that listeners had to detect a discontinuity between the two vowel halves. These two stimulus pairs were consistently marked by all listeners. The first stimulus pair always received a score of 0, which indicates that the listeners did not detect a difference between modified and reference stimulus. The second stimulus pair always received a bad score (two listeners responded with a score of 1 and four listeners responded with a score of 2), which indicates that there was an obvious discontinuity detected in the modified stimulus. All listeners agreed that judging small perceived differences was a difficult task, but after the initial training they were able to make consistent decisions. The complexity of this task warrants the use of expert listeners.

Table III presents the correlation between the subjects and the average score. The correlation between the average score and each of the subjects is quite high, between 0.49 and 0.66. The correlations between subjects are expectedly smaller, but uniformly positive.

Table IV lists the correlations between the objective distance measures. It is interesting to note that there is a high correlation between the Euclidean formant frequency distance and the SKL and MFCC distances. In addition, there is a high correlation between the MFCC distance and the overall energy distance. The Euclidean formant bandwidth distance does not correlate well with other measures.

Fig. 2 shows the correlations between the objective distance measures and the average subject score obtained from the perceptual experiment. The highest correlation is obtained between the average score and the Euclidean formant frequency distance. The SKL distance is a close second.

Table V provides the results for a linear regression carried out on the average subject score, using all of the objective distances as predictors. The variance explained R^2 was 0.23 and the correlation R was 0.48, indicating a substantial fraction of unexplained variance. The formant frequency contributed most to predicting the average subject score, followed by the SKL distance and the formant bandwidths. The MFCC and overal energy distances did not provide a significant contribution. Performing linear regression for each vowel separately revealed that the correlation was particularly low for the vowel /i:/ ($R^2 = 0.10$, $R = 0.33$), but substantial correlations were found for the remaining two vowels, with $R = 0.60$ for /@/ and $R = 0.69$ for /u/. The regression analysis further revealed that the formant frequency distance explained most of the variance for each of the three vowels. The MFCC distance was a close second for the /@/, but for the /i:/ it came in third behind the SKL distance, and for the /u/ it was one of the lower scoring distances. The

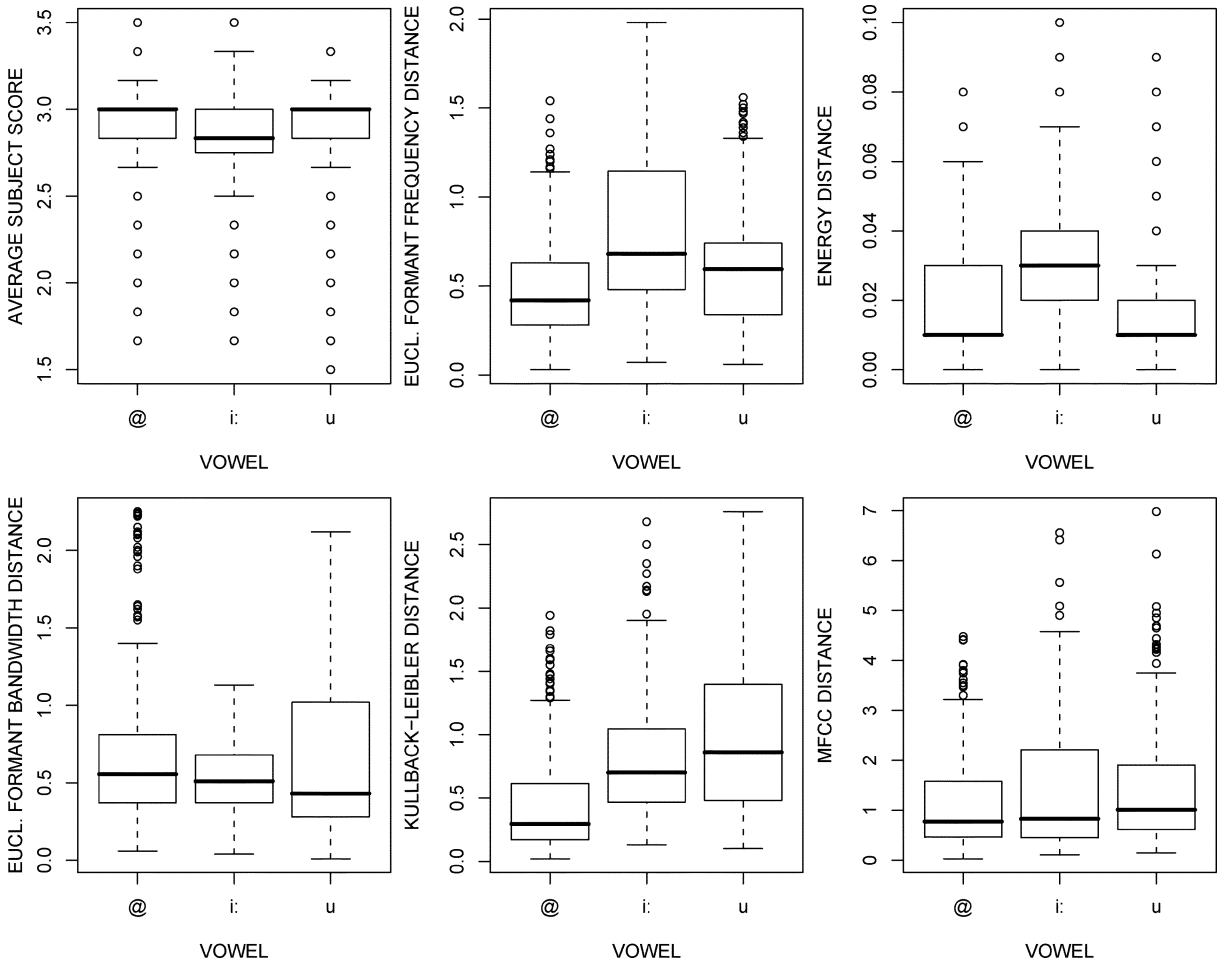

Fig. 3. Breakdown of distances per vowel.

TABLE V
RESULTS FOR LINEAR REGRESSION BETWEEN AVERAGE SUBJECT SCORE AND FIVE OBJECTIVE DISTANCE MEASURES

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-7.50	0.05	-155.83	***
D_{FF}	0.43	0.07	5.82	***
D_{SKL}	0.23	0.06	3.79	***
D_{MFCC}	0.02	0.03	0.83	
D_{E}	1.32	1.20	1.10	
D_{FB}	0.17	0.05	3.38	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '				
Residual standard error: 0.60 on 821 degrees of freedom				
Multiple R-Squared: 0.23, Adjusted R-squared: 0.23				
F-statistic: 49.55 on 5 and 821 DF, p-value: $<$ 2.2e-16				

formant bandwidth distance scored a third place for the /u/, but it was the worst performer for the other two vowels. Overall, it became apparent that none of the distances explain the variance in the /i:/ well.

Fig. 3 shows the distribution of the average subject scores and the objective distances per vowel. One of the possible explanations for the poor results of the /i:/ was that the range of distances was much lower than for the other vowels. This figure shows that this is not the case. For the formant frequency distance and the energy distance, the distances are larger than for the other two vowels. One possible explanation for the larger formant frequency distance could be that the first formant in the /i:/ is very close to the F_0 of the female speaker, making it harder to measure reliably.

III. FORMANT RESYNTHESIS EXPERIMENT

This section describes the second perceptual experiment, which serves to answer the main question of this paper. Which acoustic dimension has the largest influence on perceived discontinuity? To this end, we resynthesized words from our database using a formant synthesizer and modified one dimension at a time.

A. Procedure

The implementation of the formant resynthesis system is similar to that of a Klatt synthesizer [19]. It uses manually adjusted global values for the source parameters TL (spectral tilt, implemented as a first-order filter), and OQ (open quotient). The formant synthesizer is implemented in Matlab [20]. A function was added that computes the overall energy contour and allows it to be changed independently from the formants and bandwidths. The resulting speech still sounds quite natural because the spectrum above 4 kHz is retained. Prior to the first perceptual test, a small MOS test was conducted to determine how well the quality of the formant resynthesis compared to natural speech. In addition to the natural speech and the formant resynthesized speech, coded versions of the natural speech were included. The

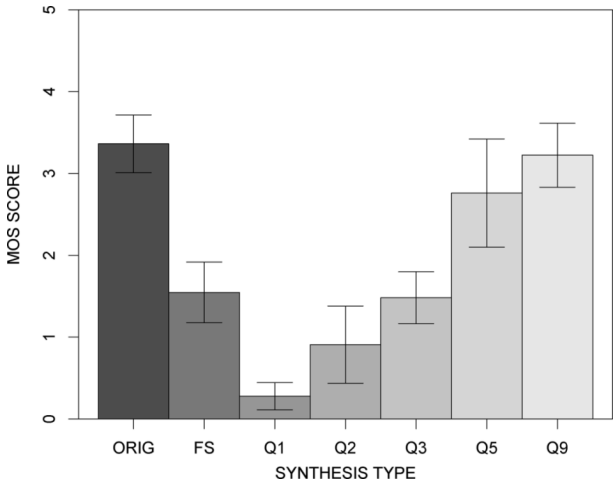

Fig. 4. Results of MOS test. Seven different versions were presented to the listener: Orig = original, FS = formant synthesis, Q1-Q9=coded with quality level 1, 2, 3, 5, or 9.

speech was coded using *Speex* [21] with different quality settings. Because there was no audible difference between quality settings 5 thru 9, the settings 6, 7, and 8 were not included. The results are presented in Fig. 4. They indicate that the formant resynthesized speech is slightly better than the coded version with quality setting 3, which corresponds to a constant bit rate conversion with a setting of 8000 b/s (bits per second). Speech encoded with this quality setting will sometimes have noticeable artefacts or noise. In comparison, quality setting 2 is said to contain very noticeable noise and artefacts, and quality setting 4 contains artefacts that can usually only be detected over high quality headphones. The reason that the MOS scores for all conditions are fairly low is that the listeners were not given an anchor point.

B. Experimental Setup

From the results of the previous experiment, we selected 20 cases with the best average subject score and 60 cases with the worst average subject score. This set of 80 stimuli was then resynthesized using formant synthesis in three conditions:

1) only the formants are changed;
2) only the bandwidths are changed;
3) only the energies are changed.

Note that this experiment did not involve concatenative synthesis. We simply took the source signal and replaced the formant values (or bandwidths or energy) of the first half of the vowel by the formant values (or bandwidths or energy) of the target stimulus and then we resynthesized the source signal with these new values using the formant synthesizer. The formant synthesis generally produced fairly good quality speech. However, the formant synthesizer only changed sonorant portions of the signal, which could lead to strange transitions between natural and resynthesized speech. In addition, the formant synthesizer sometimes had trouble resynthesizing nasals and nasalized portions adequately. Because of these considerations, we decided to remove the leading and trailing consonants from the stimuli and only present the vowel portions of each stimulus to the listener. A fade-in/-out filter was applied to the vowel edges to create smooth transitions. The reference stimulus was also formant resynthesized, but instead of inserting the formant, bandwidth or energy tracks from the target signal, the original tracks from the source signal were inserted. The task was the same as in the previous experiment. For each of the 320 stimuli, listeners had to indicate on a five-point scale which stimulus sounded better. This time 12 subjects were used, four expert subjects who had also participated in the first experiment, and eight subjects from other departments. Each block in the experiment was preceded by six examples which were not included in the analysis. One stimulus was repeated five times at random positions in the experiment for reliability testing. The experiment was supplemented with two stimuli for validity testing, one in which the modified stimulus was identical to the reference stimulus and one in which the formant frequencies of the first half of the vowel came from a different vowel than the second half. This served the same purpose as in the first experiment, the discrepancies between the formant tracks of the first and second part of the vowel were so extreme that a discontinuity was definitely present.

C. Results

The subjects were unanimous in their scores for the validity test stimuli. All of them rated the identical pair as 0 and other pair in which the formant frequencies came from a different vowel as 2. The stimulus used for reliability testing received less consistent scores, ranging from -1 to 2. We do not have an explanation for this lack of consistency. The subjects' scores were combined into an average subject score in the same manner as in the first experiment. The correlations between subjects is greater than in the first experiment, ranging between 0.35 and 0.6, which shows us that subjects were able to reliably detect discontinuities in these short segments. This can be explained by the fact that there are less sources of variability in the signals that cause the audible discontinuities and that stimuli were selected that have a substantial range in quality.

The main goal of this experiment is to determine which type of modification contributed most to the perception of discontinuity in this diphone database. This can be determined by measuring the separation in the subject scores between the 20 good cases and the 60 bad cases present for each type of modification. The hypothesis is that if a particular type of modification has a larger impact on the discontinuity perceptions, the CMOS score will be higher, i.e., there is a greater preference for the modified version over the test version. Therefore, the questions to ask are, is it the case that the 60 bad cases have consistently higher scores than the 20 good cases? And are there significant differences between the three different types of modification? This separation in scores can be computed using the robust rank order test [22], [23], chosen because it is more robust than the nonparametric Wilcoxon–Mann–Whitney test. Table VI shows the robust rank order test statistic $\hat{U}$ for each of the twelve subjects in each of the three modification conditions (formant frequencies FF, formant bandwidths FB, and overall energy E). As can be seen, the $\hat{U}$ value was generally highest for the condition where only the formant frequencies were altered, second highest for the energy modification and lowest for the formant bandwidth modifications. To test the significance of these differences, we per-

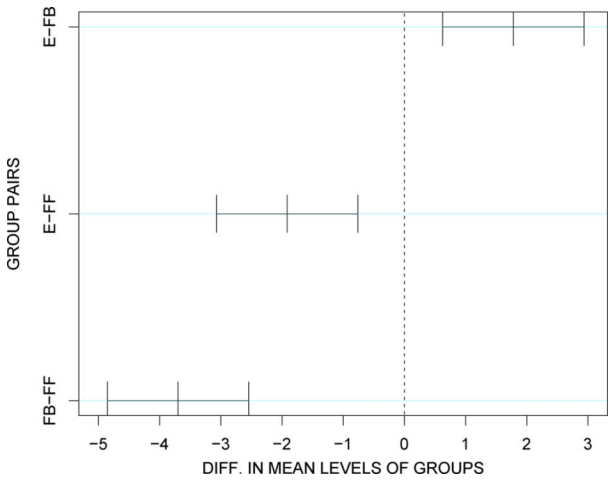

Fig. 5. Results of Tukey's HSD test. These numbers are for the 95% family-wise confidence level.

TABLE VI
$\grave{U}$ VALUES OF ROBUST RANK ORDER STATISTICS PER SUBJECTFOR EACH MODIFICATION CONDITION

	FF	FB	E
subj 1	2.95	-0.29	1.21
subj 2	4.49	0.33	2.07
subj 3	2.19	0	2.77
subj 4	3.75	0.54	1.47
subj 5	6.52	-0.06	1.19
subj 6	4.08	-0.50	0.96
subj 7	3.92	-0.15	1.48
subj 8	3.32	-0.51	2.03
subj 9	3.93	2.41	1.99
subj 10	6.22	0.03	1.28
subj 11	1.62	1.80	4.49
subj 12	4.18	-0.77	3.28

TABLE VII
ORDER OF IMPORTANCE OF DIMENSIONS OF VARIABILITY AS DETERMINED BY RANK ORDER STATISTICS

1. Formant frequency
2. Overall energy
3. Formant bandwidth

formed the Hotelling T^2 test [24]. This is a nonparametric test that provides a more reliable result than an ANOVA because it does not make any assumptions about the variance in each of the three groups. The differences turned out to be highly significant ($\mathrm{T}^2 = 340.7, \mathrm{F}(2, 33) = 92.9, p = 4.32\mathrm{e} - 07$).

The differences between each pair of groups is also significant. We performed Tukey's Honestly Significant Different (HSD) post-hoc comparison [25]. Fig. 5 confirms the results, showing that the difference between formant frequency modifications (FF) and formant bandwidth modification (FB) are largest, followed by differences between formant frequency modifications and energy modifications (E). The differences between FB and E are also significantly different from zero. The result is summarized in Table VII.

IV. CONCLUSION

This study has reported on two perceptual tests performed to dissect the causes for audible discontinuities in concatenative synthesis. The tests were designed meticulously to obtain the most reliable results. The first experiment has shown that a pairwise comparison yields subjective results that are consistent with a average subject score. The correlation between the average subject score and several objective distance measures showed that the highest correlation was obtained by the Euclidean formant frequency distance. The symmetrical Kullback–Leibler distance and the Euclidean MFCC distance also provided a good correlation, but these two measures are not independent of the formant frequency distance. When applying linear regression, the unique contributions of each measure are highlighted and the formant frequency distance proved to contribute most to explaining the overall variance. However, the effect size was different for each vowel. The best solution is therefore not to rely on a single objective distance measure, but to combine several measures employing linear regression.

The second experiment has shown that the correlation between the average subject score and the individual subjects and the correlation among subjects is even higher when only one type of mismatch is present. By computing robust rank orders we have shown that the subject scores are significantly different for each type of modification, with modifications in formant frequencies having the largest impact on the subject scores. The energy modification came second and the formant bandwidth modification came in last.

V. DISCUSSION

Because this study was restricted to speech units from a diphone inventory, the discontinuities were for the most part caused by phonetic variability caused by coarticulation. This is the main reason that we chose to investigate the contribution of formant frequencies, formant bandwidths and overall energy to audible discontinuities. In unit selection systems, where units vary not only in their phonetic context but also in their prosodic context, other factors such as voice quality and glottalization will become more relevant. Apart from looking at the overall energy in this study, we have also considered a four-band energy measure which breaks the energy up into four broad frequency bands to reflect spectral balance variability [26], [27]. Even though this spectral balance measure did not correlate very well with our average subject score, previous research has shown that this measure is important in cases where prosodic contexts vary systematically, as is the case in unit selection synthesis.

The second experiment in which the vowels were resynthesized using formant resynthesis has shown that formant synthesis is a viable option which can result in high-quality speech, when the upper part of the spectrum is maintained. We are currently in the process of creating a new synthesis system that builds on the formant synthesizer discussed in this study. This new synthesizer allows systematic changes to spectral balance, formant frequencies and bandwidths. The research involved in building this synthesizer serves to answer the most important question: how can we achieve high-quality synthetic speech with smooth transitions between units? This will involve investigating different smoothing methods to smooth formant tracks, but it is expected that a more sophisticated modeling technique will be necessary to generate acceptable formant trajectories to be used in the synthesis process.

REFERENCES

[1] M. Balestri, A. Pacchiotti, S. Quazza, P. L. Salza, and S. Sandri, "Choose the best to modify the least: A new generation concatenative synthesis system," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999, pp. 2291–2294.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, 1996, pp. 373–376.

[3] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers," in *Proc. 4th ISCA Speech Synthesis Workshop*, Pitlochry, U.K., 2001, pp. 123–126.

[4] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 39–51, Jan. 2001.

[5] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, Salt Lake City, UT, 2001, pp. 837–840.

[6] A. Syrdal, "Phonetic effects on listener detection of vowel concatenation," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 979–982.

[7] A. Syrdal and A. Conkie, "Data-driven perceptually based join costs," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 49–54.

[8] ——, "Perceptually based data-driven join costs: Comparing join types," in *Proc. Eurospeech'05*, Lisbon, Portugal, 2005, pp. 2813–2816.

[9] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds. Upper Saddle River, NJ: Prentice-Hall, 2004, pp. 35–62.

[10] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP'98)*, Sydney, Australia, 1998, vol. 6, pp. 2747–2750.

[11] J. Hieronymus, "ASCII phonetic symbols for the world's languages: Worldbet," Tech. Memo, AT&T Bell Laboratories, Murray Hill, NJ, 1994.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3-4, pp. 187–207, 1999.

[13] ITU P.800, 1996, "Methods for subjective determination of transmission quality," Int. Telecomm. Union (ITU). Rec. P.800, 2005 [Online]. Available: http://www.itu.int

[14] T. Veenker, "WWStim: A CGI script for presenting web-based questionnaires and experiments," 2005 [Online]. Available: http://www.let.uu.nl/ Theo.Veenker/personal/projects/wwstim/doc/en/

[15] *The Snack Sound Toolkit*, Royal Inst. Technol., Oct. 2005 [Online]. Available: http://www.speech.kth.se/snack/

[16] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, p. 248, 1961.

[17] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 97–100, 1990.

[18] B. Moore, "Parallels between frequency selectivity measured physophysically and in cochlear mechanics," *Scand. Audiol., Supplement*, vol. 25, pp. 139–152, 1986.

[19] J. Allen, S. Hunnicut, and D. Klatt, *Text-to-Speech: The MITalk System*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[20] A. Kain, X. Niu, J. P. Hosom, Q. Miao, and J. P. H. van Santen, "Formant re-synthesis of dysarthric speech," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 25–30.

[21] J. M. Valin, "The Speex Codec Manual," 2005 [Online]. Available: http://www.speex.org/manual.ps, 1.0

[22] N. Feltovich, "Critical values for the robust rank order test," *Commun. Statist.*, vol. 34, no. 3, pp. 525–548, 2005.

[23] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. New York: McGraw-Hill, 1988, pp. 206–215.

[24] D. F. Morrison, "Tests of hypotheses on means," in *Multivariate Statistical Methods*, 4th ed. New York: McGraw-Hill, 1990, ch. 4.

[25] W. L. Hays, "Comparisons among means," in *Statistics*. Orlando, FL: Holt, Rinehart, and Wilson, 1988, ch. 11, pp. 418–421.

[26] J. P. H. van Santen and X. Niu, "Prediction and synthesis of prosodic effects on spectral balance of vowels," in *Proc. 4th IEEE Workshop Speech Synthesis*, Santa Monica, CA, 2002, pp. 147–150.

[27] Q. Miao, X. Niu, E. Klabbers, and J. P. H. van Santen, "Effects of prosodic factors on spectral balance: Analysis and synthesis," in *Proc. 3rd Int. Conf. Speech Prosody*, Dresden, Germany, 2006, (CDROM).

Esther Klabbers received the M.A. degree in language and computer science from the Department of Language and Speech, University of Nijmegen, The Netherlands, in 1995 and the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2000.

Since 2001, she has been a Senior Research Associate at the Center for Spoken Language Understanding (CSLU), OGI School of Science Engineering, Oregon Health and Science University, Beaverton, OR. Her research interests include speech synthesis, prosody modeling, and perceptual testing.

Jan P. H. van Santen (M'00) received the Ph.D. degree in mathematical psychology from the University of Michigan, Ann Arbor, in 1979.

He is the Director of the Center for Spoken Language Understanding and a Professor in the Biomedical Engineering (BME) and Computer Science and Electrical Engineering (CSEE) departments at the OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR. He is currently also the Department Chair of CSEE. He is also founder and director of Biospeech, Inc., a spin-off company that focuses on the creation of cost-effective products for communication disorders using advanced speech and language technologies. His research focuses on mathematical modeling of prosody, signal processing, and computational linguistics. A key growing focus of his work and of the Center is on basic and applied speech and language technology research for communication disorders. He was a Member of the Technical Staff at Bell Laboratories, Lucent Technologies, Murray Hill, NJ, from 1984 to 2000, an Associate Research Scientist at New York University from 1981 to 1984, and a Postdoctoral Fellow at Bell Laboratories from 1979 to 1981.

Alexander Kain received the B.A. degree in computer science and the B.A. degree in mathematics from Rockford College, Rockford, IL, in 1995 and the Ph.D. degree in computer science from the OGI School of Science and Engineering, Oregon Health and Science University, Beaverton, OR, in 2001.

From 2001 to 2005, he was a Senior Research Associate at the Center for Spoken Language Understanding (CSLU), OGI School of Science and Engineering, and a Lead Speech Synthesis Technologist at Sensory, Inc., Portland, OR. He is currently an Assistant Scientist at CSLU and a Research Scientist at Biospeech, Inc. His research focuses on signal processing, with a focus on speech modifications for enhancing intelligibility.