Reducing Audible Spectral Discontinuities

Esther Klabbers and Raymond Veldhuis

***Abstract*—In this paper, a common problem in diphone synthesis is discussed, *viz.*, the occurrence of audible discontinuities at diphone boundaries. Informal observations show that spectral mismatch is most likely the cause of this phenomenon. We first set out to find an objective spectral measure for discontinuity. To this end, several spectral distance measures are related to the results of a listening experiment. Then, we studied the feasibility of extending the diphone database with context-sensitive diphones to reduce the occurrence of audible discontinuities. The number of additional diphones is limited by clustering consonant contexts that have a similar effect on the surrounding vowels on the basis of the best performing distance measure. A listening experiment has shown that the addition of these context-sensitive diphones significantly reduces the amount of audible discontinuities.**

***Index Terms*—Audible discontinuities, context-sensitive diphones, spectral distance measures.**

I. Introduction

ONE well-known problem with concatenative synthesis is the occurrence of audible discontinuities at concatenation points. In our database of one female speaker, this is most prominent in vowels and semi-vowels. It is due to variability in the pronunciation of these sounds which is caused by the phonetic/prosodic context.

Discontinuities are caused by mismatches in F_0, phase or spectral envelopes across concatenation points [8]. In Calipso, IPO's diphone synthesis system [29], F_0 mismatches are avoided by monotonizing the diphones before storing them in the database. Phase mismatches are avoided by using a method called *phase synthesis* for re-synthesis of the nonsense words [9]. Phase synthesis is based on accurate measurements of the mixture of periodic and noise information in speech. The input speech is analyzed pitch-synchronously like in TD-PSOLA, but the pitch periods are estimated more precisely by means of "first-harmonic filtering." This forms the basis of a Discrete Fourier Transform (DFT), providing exact amplitude and phases for all harmonics. It uses overlap-and-add over two pitch periods. The signal is reconstructed by means of an amplitude and a phase value for each harmonic. Because the harmonics are added with coherent phases, phase mismatches are avoided.

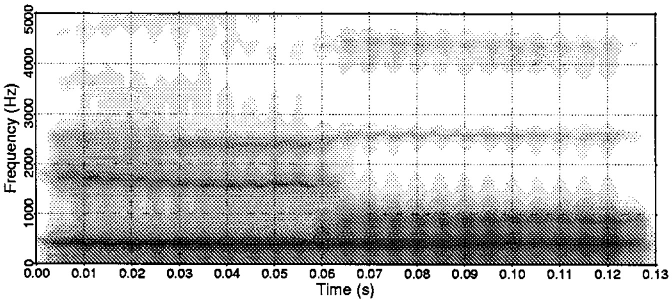

Fig. 1. Spectrogram for the vowel /u/ in the synthesized Dutch word *doek*. A considerable mismatch in F_2 between the left and right half of the phoneme is visible. The sudden transition at the concatenation point causes an audible discontinuity.

Spectral mismatch is still a major problem, though. As an example, consider Fig. 1, which shows the spectrogram for the vowel /u/ in the synthesized Dutch word *doek* (consisting of the diphones /du/ + /uk/). It reveals a considerable mismatch in F_2 between the first and second half of the phoneme. An audible discontinuity was clearly present. This, along with other informal observations, suggests that spectral mismatch is the main cause for the occurrence of audible discontinuities. Please note that pitch and phase mismatches have already been eliminated.

In order to solve the problem of spectral mismatch, several solutions have been proposed. One approach is to use larger units such as demi-syllables or triphones. However, this does not solve the problem, as discontinuities continue to occur albeit less frequently. Moreover, the database size increases drastically. As [24] point out, in American English assuming a 43-phone alphabet, at least 70 000 of the theoretical maximum of 79 507 triphones actually occur in the language. Even when incorporating all these units, smooth joins are not guaranteed as all triphones can occur in different contexts with strong coarticulatory effects that can even span word boundaries.

Another approach is to vary the location of the cutting points in the nonsense words dependent on the context [6]. This calls for a spectral distance measure that correctly represents the amount of spectral mismatch. This method is based on the assumption that the short-term spectral envelopes are not constant over time, resulting for instance in nonflat formant trajectories, and that therefore appropriate cutting points can be found. However, Fig. 1, along with many other observations in our database, shows that formant trajectories can be fairly flat throughout a vowel when they are embedded in nonsense words of the type C@CVC@ (C = consonant, V = vowel, @ = schwa).

A third approach is to perform smoothing by means of waveform interpolation, spectral-envelope interpolation or formant trajectory smoothing. It requires specific signal representations that allow these types of operations. The disadvantage of formants as a representation is that they are very difficult to estimate reliably. Waveform and spectral envelope interpolation have the disadvantage that smooth transitions are often achieved at the expense of naturalness [8]. Examples of signal representations that allow waveform interpolation are multi-band resyn-

Manuscript received June 26, 2000; revised September 22, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Juergen Schroeter.

The authors are with IPO, Center for User-System Interaction, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands (e-mail: e.a.m.klabbers@tue.nl; r.n.j.veldhuis@tue.nl).

Publisher Item Identifier S 1063-6676(01)00333-9.

TABLE I
COMPOSITION OF MATERIAL FOR THE PERCEPTUAL EXPERIMENT; THE TOTAL NUMBER OF $C_l VC_r$ STIMULI IS 2415 23 $C_l \times$ 5 V $\times$ 21 C_r

C_l:	p	t	k	b	d	g	f	s	x	S	v	z	G
	Z	m	n	l	L	r	j	w	c	h			
V:	a	A	i	I	u								
C_r:	p	t	k	b	d	g	f	s	x	S	v	z	G
	Z	m	n	l	L	r	j	w					

TABLE II
PERCENTAGE OF PERCEIVED DISCONTINUITIES PER VOWEL. THE PERCENTAGES ARE COMPUTED FROM THE SUM OF THE MAJORITY SCORES

Vowel	Percentage of perceived discontinuities	Number of observations
/a:/	17.1%	474
/i/	43.1%	445
/A/	52.1%	468
/I/	55.5%	449
/u/	73.9%	448

thesis overlap-and-add (MBROLA) [8] and harmonic plus noise modeling (HNM) [28]. Reference [5] present several different techniques for spectral smoothing, none of which they found really satisfactory.

A fourth approach is to include context-sensitive or specialized units in the database [24]. This implies that one knows which contexts can be clustered so as to keep the database size within bounds. Our investigation is aimed at gaining insight in this approach. In this paper, we first present a detailed analysis of the occurrence of audible discontinuities in our diphone database (Section II). The aim of this part of the study was to find an objective spectral distance measure that best predicts when discontinuities are audible. Therefore, we related the results of a perceptual experiment with several distance measures. In Section III we study the feasibility of extending the diphone database with context-sensitive diphones to reduce the occurrence of audible discontinuities. The number of additional diphones is limited by clustering similar contexts on the basis of the best performing distance measure. This approach goes toward the use of triphones, except that it requires a smaller extension of the database than when true triphones are added.

II. ANALYSIS OF THE PROBLEM

A. Perceptual Experiment

The first step in our analysis was to find out to what extent audible discontinuities occur in our diphone database. This was established via a perceptual experiment. IPO's speech-synthesis system Calipso currently uses diphones as concatenative units from a professional female speaker. They have been excised from nonsense words. For instance, consonant–vowel (CV) and vowel–consonant (VC) diphones are excised from nonsense words of the form C@CVC@. In order to reduce the data set to manageable proportions, this study was restricted to five Dutch vowels in this database, i.e., the vowels /a:/, /i/, /A/, /I/, /u/ (in SAMPA notation). The vowels /a:/, /i/, and /u/ were chosen because they cover the extremes in the vowel space. The vowels /A/ and /I/ are chosen because they are the short counterparts for /a:/ and /i/. See [3] for an overview of the Dutch phoneme inventory. A study by [13] has shown that coarticulation is speaker-specific. Therefore, it should be noted that the results presented in this paper only reflect the coarticulatory behavior of the speaker of our diphone database.

1) Material: The stimuli consisted of concatenated left C_lV and right VC_r diphones, which were excised from the nonsense words C_l@C_lVC_l@ and C_r@C_rVC_r@. The stimuli consisted of five vowel conditions in the context of all consonant pairs that can occur in C_l and C_r position (see Table I). The total number of stimuli is $23 \times 5 \times 21 = 2415$. So, for instance, the diphones /du/ and /uk/ that form the stimulus /duk/ were extracted from the nonsense words *d@dud@* and *k@kuk@*. The diphones were created using the phase synthesis technique mentioned in Section I. No spectral smoothing was applied at the boundary.

In the stimuli, the consonant portions were cut off to prevent them from influencing the perception of the diphone transition in the middle of the vowel.[1] Fading was used to smooth the transition from silence to vowel and vice versa. Because all stimuli were presented in isolation, the stimulus duration had to be long enough to be able to perceive the transition at the diphone boundary. The duration of the vowels was fixed to 130 ms with the diphone boundary located exactly in the middle of the vowel. The signal power of the second diphone was scaled to match that of the first diphone.

2) Procedure: Five participants with a background in psycho-acoustics or phonetics participated in the perceptual experiment. It was a within-subjects design meaning that each subject received all stimuli in random order. For each stimulus, the participants had to judge the transition at the diphone boundary as either smooth (0) or discontinuous (1). The experiment was divided into three hourly sessions which were held on different days, with a short break halfway through each session. The session order was different for all participants. The experiment started with a familiarization phase in which two stimuli were presented for each vowel, one being clearly smooth and the other being clearly discontinuous. The setup of this experiment results in very critical observations because 1) the vowels have been placed out of context and 2) subjects are forced to make a binary decision. This provides a more critical test than when using real speech.

3) Results: The participants found the task difficult, but felt they had been able to make consistent decisions after the familiarization phase. As a consistency check, we presented two stimuli, one clearly smooth, the other clearly discontinuous, ten times at random positions in the total stimulus list. All participants were 100% consistent in their scoring of these two stimuli. Between participants there was more variability, as some participants applied a stricter threshold than others. In order to reduce the variability between participants, a majority score was calculated, i.e., a stimulus was marked as discontinuous when four out of five listeners perceived it as such. Summing the majority scores obtained in the experiment for each of the vowels, we get the percentage of perceived discontinuities as presented in Table II.

[1] Just before running the experiment we discovered that in a few cases the influence of the first consonant was still audible in the vowel. Instead of changing the duration of all vowels and re-excising the stimuli, we decided to discard these cases, leaving 2284 stimuli to be judged in the perceptual experiment. We believe that this did not influence the results.

The results show that the number of audible discontinuities is particularly high for /u/ and comparatively low for /a:/. Our results additionally reveal a slightly better score for the long vowels /a:/ and /i/ than for the short vowels /A/, /I/ and /u/. This is partly in line with findings by [13]. They investigated speaker variability in the coarticulation of /a:/, /i/ and /u/. Their results show that the /u/ has the greatest amount of coarticulation and the /i/ has the smallest amount, closely followed by /a:/.

B. Spectral Distance Measures

The second step in our investigation was to relate the results from the perceptual experiment with several spectral distance measures in order to obtain an objective measure for predicting audible discontinuity. In speech recognition and speech coding, spectral distance measures have been widely used. In automatic speech recognition, one of the earliest studies comparing several distance measures was conducted by [10]. They investigated measures based on spectral and cepstral coefficients, log area ratios and the Itakura–Saito distance. They obtained the best performance with the root-mean-squared (rms) log spectral distance. [11] and [17] showed that using warped frequency scales (such as Mel-scale or Bark-scale) improved the performance of speech recognizers even further. The most commonly used distance measure in automatic speech recognition is the Euclidean distance between Mel-frequency cepstral coefficients (MFCC).

In speech synthesis, this Euclidean MFCC distance has also been adopted in order to select optimal units or segment diphones at the optimal cutting point (among others by [6] and [4]). The question is whether a measure used in speech recognition is equally suitable for use in speech synthesis, as it serves a different purpose. In speech recognition, the task is to classify different instances of one and the same phoneme as belonging to the same target phoneme, whereas in speech synthesis the task is to distinguish these instances when their spectra are perceptually different. Therefore, it should be investigated whether some distance measures can be found that correspond to human perception in that they are able to distinguish perceptually relevant differences in spectra [26].

An investigation that ran parallel to ours [20], [31] also aimed at performing a perceptual evaluation of distance measures in the context of speech synthesis. In their study, listeners had to judge the difference between a pair of stimuli on a scale from zero to five. One stimulus was the reference stimulus produced by a diphone synthesis system, the other stimulus was altered in that the first (c.q. second) half of the vowel phoneme was replaced by a different instance of the vowel preceded (c.q. followed) by a consonant from the same class as the original. Five feature representations were studied: FFT-based cepstra, LPC-based cepstra, line spectral frequencies (LSF), log area ratios (LAR), and a symmetrized Itakura distance. All but the FFT-based cepstra were computed from LPC coefficients. The feature representations were computed in three ways:

1) using the FFT amplitude spectrum;
2) using a perceptual spectrum (PLP, [12]);
3) using a Mel-warped spectrum.

Correlations between the average of the listeners' responses and the distance measures were computed and then combined into a population correlation using Fisher's z-transform. Correlations were not particularly high, reaching from 0.28 for the linear log area ratio to 0.50 for the linear Itakura distance. PLP and Mel-scale improved the correlations, but the improvement from PLP to Mel was not significant. Using a weighted Euclidean distance improved the linear measures, but only slightly for PLP and Mel. Delta features gave only a small increase in correlation (0.02). The best correlation was obtained for Mel cepstra with delta features (0.66), where it did not make a difference whether these were computed from FFT or LPC coefficients.

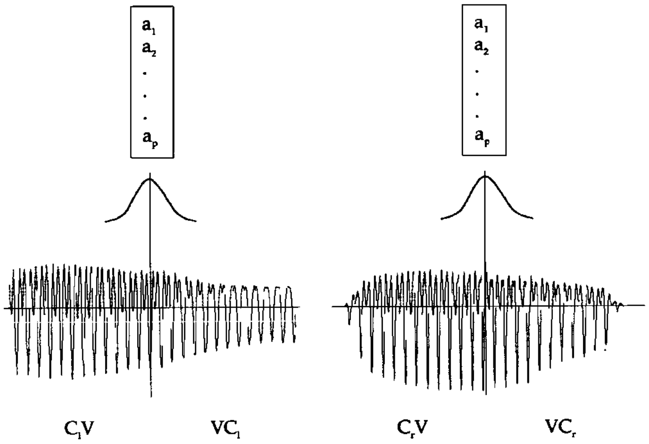

Fig. 2. Computation of LPC-coefficients $(p = 14)$ at the diphone boundary in the CVC-part of the nonsense words, using a 40-ms Hanning window.

The measures used in this paper are taken from various fields of research. They were used to determine distances between spectral envelopes across diphone boundaries. The following spectral distance measures were used. They will be explained in more detail below.

1) Euclidean distance between (F_1, F_2) pairs, or the Euclidean formant distance (D_{EFD}), which is often used in phonetics.
2) Symmetrical Kullback–Leibler distance (D_{SKL}), which originates from the field of statistics.
3) Partial loudness D_{PL}, which comes from the area of sound perception.
4) Euclidean distance between Mel-frequency cepstral coefficients (D_{MFCC}), which comes from automatic speech recognition.
5) Likelihood ratio (D_{LR}), which is used in speech coding and automatic speech recognition.
6) The mean-squared log-spectral distance (D_{MSLSD}), which also comes from automatic speech recognition.

All spectral distances excepting the Euclidean formant distance, were calculated from LPC-spectral envelopes. The sampling frequency of the speech was 16 kHz. First, two sets of LPC coefficients $a_1, \cdots, a_p$ $(p = 14)$ were computed from Hanning-windowed signal segments of 40 ms symmetrically positioned around the diphone boundary [see Fig. 2]. One set is measured at the right diphone boundary of the C$_l$V diphone in the nonsense word C@C$_l$VC$_l$@, which also produces the diphone VC$_l$. The other set of LPC coefficients is measured at the left diphone boundary of the VC$_r$ diphone in the nonsense word C@C$_r$VC$_r$@, which also produces the diphone C$_r$V. A pre-em-

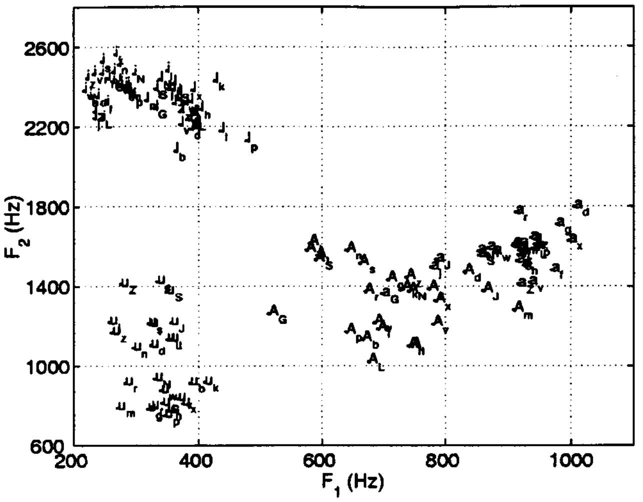

Fig. 3. Vowel space in terms of F_1 and F_1 for the five vowels /a:/, /A/, /i/, /I/, and /u/ in different (symmetric consonantal contexts as indicated by the subscripts).

phasis filter with a transfer function $H(z) = 1 - 0.95z^{-1}$ was applied. From those LPC coefficients, two power spectra were computed at sufficiently many equidistant frequency points (in our case 512), which were arranged in a vector. These two vectors were normalized by dividing them by the sum of the elements. This gives two power-normalized spectral envelopes $P(\omega)$ of C$_l$V and $Q(\omega)$ of VC$_r$. Equation (1) displays the computation of $P(\omega)$ from the LPC coefficients, where K = constant, such that $\int P(\omega)\, d\omega = 1$. The distance measures will be discussed in more detail in the following:

$$P(\omega) = \frac{K}{|1 + a_1 e^{-j\omega} + a_2 e^{-2j\omega} + \cdots + a_p e^{-pj\omega}|^2}. \quad (1)$$

1) Euclidean Formant Distance: In phonetics, it is quite common to describe coarticulation in terms of changes in the formants F_1 and F_2 [13], [24]. In this investigation, the formants were measured automatically at the diphone boundary in the vowel using Praat [2]. These were verified and corrected when necessary. Close inspection of the stimuli reveals that most formant trajectories are fairly stationary throughout the vowel, except when the surrounding consonants are the alveolars /j/, /J/, /c/, /S/ and /Z/.

Fig. 3 displays the F_1 and F_2 values for the five vowels measured in our diphone database at the indicated locations. It shows that the /a:/, /i/ and /I/ have small variations, whereas the /A/ and /u/ seem to be affected to a greater extent by their surrounding consonants. For the /u/, differences in F_2 are considerable. They can be as large as 700 Hz. For the /A/, differences in F_1 are also considerable. The /i/ and /I/ are very close to each other in the vowel space. All formant frequencies were transformed to a Mel-scale which is more in line with the hearing process than a linear frequency scale. For the Mel transformation, we used (2) [21]

$$m = 2595 \log\left(1 + \frac{f}{700}\right). \quad (2)$$

The Euclidean formant distance is calculated by

$$D_{\text{EFD}}(l, r) = \sqrt{(F_{1,l} - F_{1,r})^2 + (F_{2,l} - F_{2,r})^2}. \quad (3)$$

2) Symmetrical Kullback–Leibler Distance: The Kullback–Leibler (KL) distance or *relative entropy* is a measure taken from statistics [18], where it is used to compute the distance between two probability distributions. Here, it is calculated from the two power-normalized spectral envelopes $P(\omega)$ and $Q(\omega)$ that were explained earlier. The original asymmetrical definition of the KL distance is changed into a symmetrical version. The main reason is that $D_{\text{SKL}}(P(\omega), Q(\omega))$ then equals $D_{\text{SKL}}(Q(\omega), P(\omega))$, such that the measure does not depend on the order of the arguments supplied.

The SKL distance has the important property that it emphasizes differences in spectral regions with high energy more than differences in spectral regions with low energy. Thus, spectral peaks are emphasized more than valleys between the peaks and low frequencies are emphasized more than high frequencies, due to the 6 dB/octave declination in spectral energy that results from the combination of the damping of the high-frequency components in the signal (−12 dB/octave) and the radiation at the mouth (+6 dB/octave). The definition for $D_{\text{SKL}}(P, Q)$ is given by

$$D_{\text{SKL}}(P, Q) = \int (P(\omega) - Q(\omega)) \log\left(\frac{P(\omega)}{Q(\omega)}\right) d\omega. \quad (4)$$

3) Partial Loudness: The partial loudness comes from the area of sound perception [23]. In a study by [7], partial loudness was shown to be a reasonably good predictor for audibility discrimination thresholds. Therefore, it was decided to include this measure to see how well it would predict audible discontinuity. The partial loudness of a signal is the loudness of the signal when presented in a background sound. The background sound generally reduces the loudness of the signal. This effect is called partial masking. The loudness of a signal in a background sound is therefore called partial loudness. In this study, the excitation patterns E_l and E_r at the C$_l$V and VC$_r$ diphone boundaries were computed. These excitation patterns are decomposed into excitation patterns representing the background sound $(\min(E_l, E_r))$, the total sound $(\max(E_l, E_r))$ and the absolute difference $(|E_l - E_r|)$. The resulting excitation patterns are then fed into Moore's partial loudness model.

4) Mel-Frequency Cepstral Coefficients: In the field of speech recognition, Mel-frequency cepstral coefficients (MFCC) are currently the most commonly used type of signal representation. They provide a successful and efficient way to represent the signal for the purpose of recognition. The MFCC coefficients were computed as described in [25, ch. 4], except that samples of the LPC power spectrum [see definition for $P(\omega)$ above] were used instead of the squared magnitudes of the DFT spectrum. Equation (5) computes the Euclidean distance between cepstral coefficients.

The cepstral coefficients can be interpreted in the following way. The c_0 coefficient represents the average energy in the speech frame. It is not included in the distance measure. c_1 reflects the energy balance between low and high frequencies (or *spectral tilt*), higher values indicating sonorants and low values suggesting frication. Higher order cepstral coefficients reflect increasing spectral detail, but no simple relationship exists between these parameters and formants [14]. In this study, the order p is set to 22

$$D_{\text{MFCC}} = \sum_{k=1}^{p} (c_{k,l} - c_{k,r})^2. \tag{5}$$

Delta features are estimations of the time derivatives of the static features, thus capturing more of the speech dynamics. Because our stimuli consist of two parts that are both fairly stationary, the addition of delta features will not make much difference to the result. In the study presented in [31] and [20] which used less restricted nonsense words, the added effect of delta features was also negligible. The correlation between the distance measure and the perceptual judgements increased by just 0.02. Therefore, it was decided not to include delta features in our investigation.

5) Likelihood Ratio: The likelihood ratio or *Itakura distance*, is a measure of spectral similarity between two LPC vectors a_L and a_R, which represent the left and right diphones [15]. It indicates how well the analysis filter of the left diphone matches that of the right diphone. It is defined in terms of an autocorrelation function. V_R represents the signal autocorrelation matrix that gave rise to a_R. The likelihood ratio is computed with (6) taken from [25, ch. 4]

$$D_{\text{LR}}(a_L, a_R) = \frac{a_L' V_R a_L}{a_R' V_R a_R} - 1. \tag{6}$$

6) Mean-Squared Log Spectral Distance: The mean-squared log spectral distance (MS LSD) is derived from the log spectral distance presented in [25, ch. 4], and is computed by (7). It is similar to the rms log spectral distance that performed best in the automatic speech recognition experiments by [22]. By taking the logarithmic differences between $P(\omega)$ and $Q(\omega)$, it is expected to better reflect the hearing process

$$D_{\text{MSLSD}} = \int (\log P(\omega) - \log Q(\omega))^2 \, d\omega. \tag{7}$$

C. Relating the Results

In order to find out how well the different spectral distance measures could predict audible discontinuities, it was decided to use receiver operator characteristic curves [19], coming from signal detection theory. Because the relation between the spectral distances and the subjects' scores was not linear, a statistical correlation could not be performed. The procedure works as follows. In order to relate the measures to the scores of the listeners, two probability density functions, $p(D|0)$ and $p(D|1)$, are estimated from the data, representing the probability of a spectral distance (D) given that the transition was marked by the listeners as continuous (0) or discontinuous (1), respectively. For a certain threshold β, the probability of a *false alarm*, the case that a transition is wrongly classified as discontinuous, is $P_F(\beta)$ [see (8)] and the probability of a *hit*, the correct detection of a discontinuity, is $P_D(\beta)$ [see (9)]

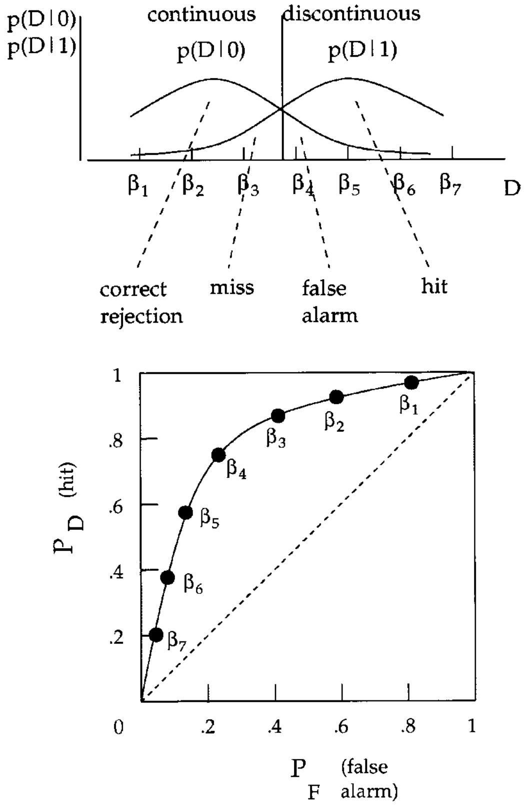

Fig. 4. Principle of receiver operating characteristic curves. The top panel displays two probability density functions $p(D|0)$ and $p(D|1)$, for the distribution of distances given that the participants have judged a stimulus as smooth and discontinuous, respectively.

$$P_F(\beta) = \int_{\beta}^{\infty} p(D|0) \, dD \tag{8}$$

$$P_D(\beta) = \int_{\beta}^{\infty} p(D|1) \, dD. \tag{9}$$

The probability of a *miss*, the case that a discontinuity goes undetected, is $1 - P_D$ and the probability of a *correct rejection*, the case that a transition is rightly classified as being smooth, is $1 - P_F$. Since these are directly derivable from the hit and false alarm probabilities, they are not relevant here.

A plot of pairs $(P_D(\beta), P_F(\beta))$ for all values of β constitutes a receiver operating characteristic (ROC) curve. See Fig. 4 for a schematic representation of ROC curves. In this experiment, we assumed that the participants were correct in their judgements. The question is then how well a spectral distance measure can predict the participants' judgements. ROC curves are always upward concave. The straight line represents the chance level meaning that a measure gives no information. The further the curve extends to the upper left corner, the better the measure serves as a predictor. This indicates that the two probability density functions $p(D|0)$ and $p(D|1)$ are moving away from each other, thus increasing the hit rate and decreasing the false alarm rate. We do not have to decide on an appropriate threshold in this study, since the purpose of the analysis is solely to determine the best performing distance measure relative to the other measures.

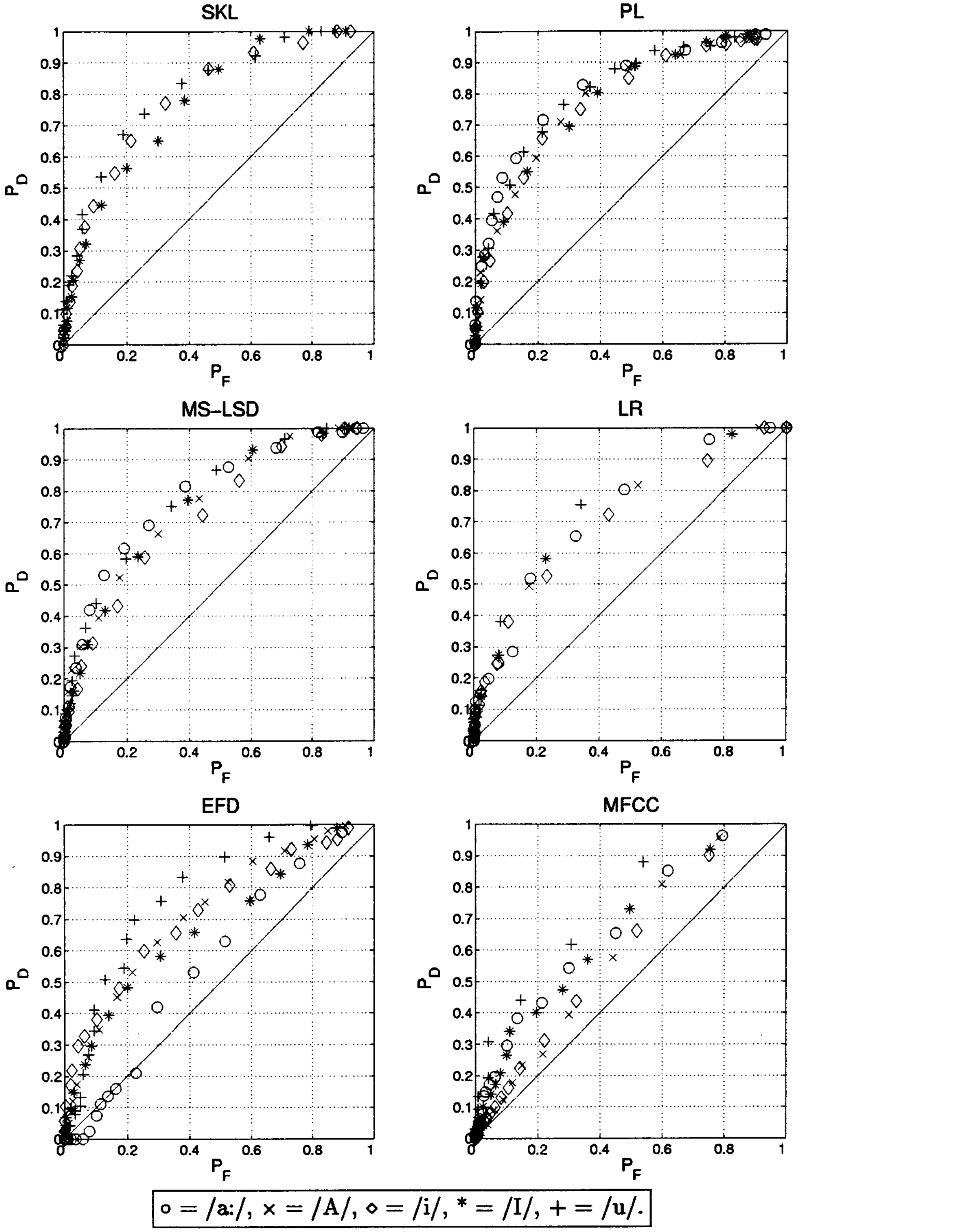

Fig. 5. ROC curves grouped per vowel for the measures SKL, PL, MSLSD, LR, EFD, and MFCC.

D. Results

Fig. 5 displays five ROC curves per distance measure, one for each vowel. Their inspection leads to a number of interesting observations. First, it can be observed that the SKL and PL distances perform equally well for all vowels, whereas the divergence between vowels is greater for the other measures.

Second, it can be seen that the Euclidean Formant Distance performs well for /u/, poorly for /a:/ and moderately well for the other vowels. This is understandable as Fig. 4 showed that /u/ had the largest degree of variation in the second formant. Extending the distance with F_3 and F_4 did not enhance the measure in any way. It was decided not to include formant bandwidths in the distance measure. As listed in [25] the just-noticeable-dif-

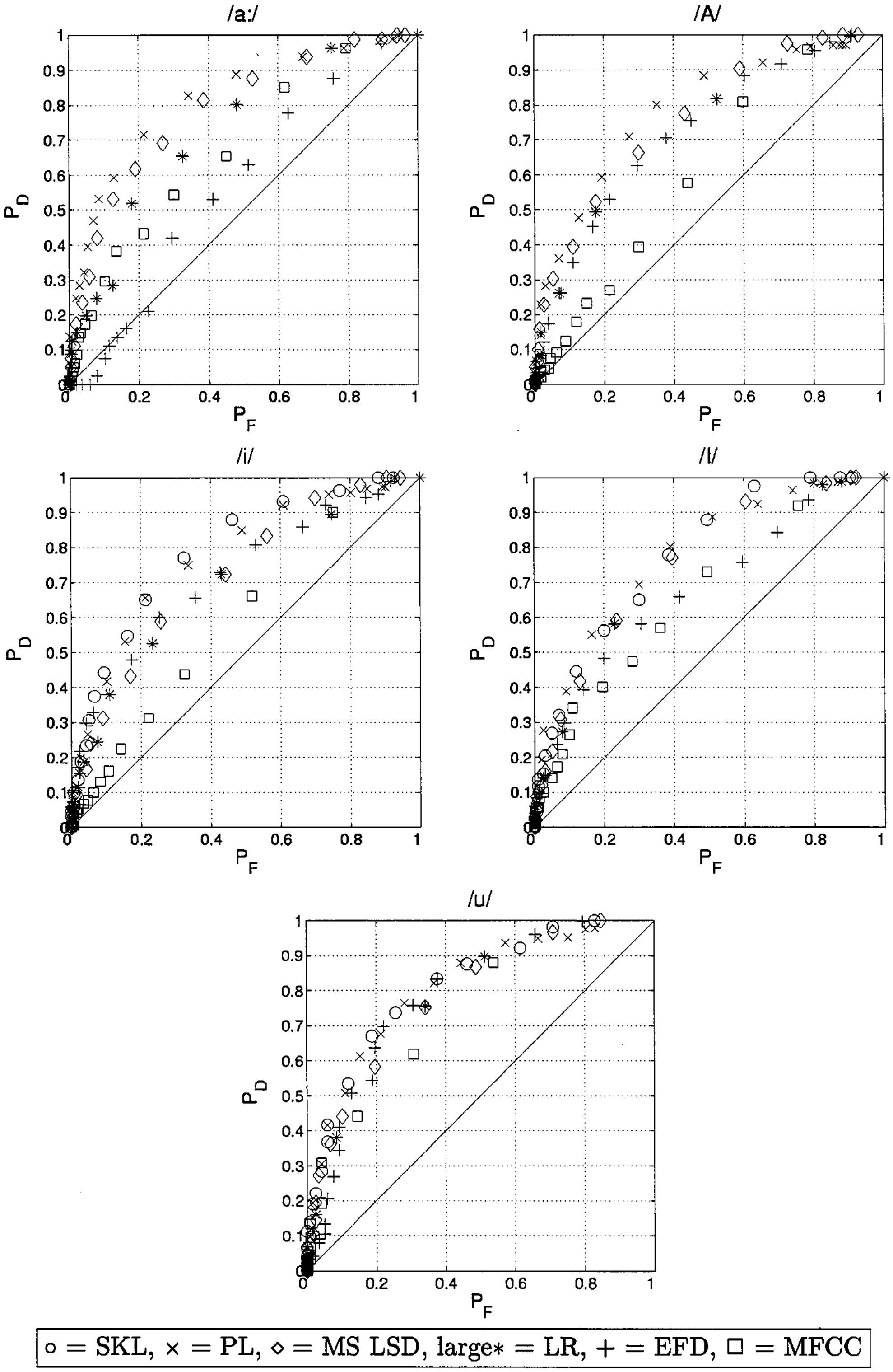

Fig. 6. ROC curves grouped per measure for the vowels /a:/, /A/, /i/, /I/, and /u/.

ference (jnd) for formant bandwidths is much larger (20–40%) than for formant frequencies (3–5%). It was therefore not expected to add much to the performance of the EFD.

Fig. 6 displays six ROC curves per vowel, one for each distance measure. Here, it can clearly be seen that all spectral distance measures perform almost equally well for the /u/, whereas the divergence between them is much larger for the other vowels.

The Euclidean distance between Mel-frequency cepstral coefficients is consistently among the worst predictors of audible discontinuity. In some cases, it is barely above chance level. This is a surprising result, because until now it was a commonly used distance measure for this task. However, its bad performance is understandable when we consider that this measure is almost standard as a distance measure in automatic speech recognition, where its task is to group together different allophones of a phoneme instead of distinguishing them when their spectral characteristics lead to perceptual differences.

Two measures are always positioned in the most upper left corner, meaning that they always performed best, the partial loudness (PL) and the symmetrical Kullback–Leibler (SKL) distance. Apparently, they correlate well with human perception. The fact that the relatively simple Euclidean formant distance only performs well for /u/ but not as well for other vowels, indicates that coarticulation affects the vowels differently. It could be that coarticulation has a larger effect on the formants for /u/, whereas in the other vowels it manifests itself in other ways, e.g., via changes in spectral tilt. Additionally, the fact that the SKL and PL distances *are* good predictors of audible discontinuity for all vowels, shows that there is systematic variation in the signal due to coarticulation that these measures capture. Because the SKL distance is much easier and faster to compute than the PL distance, it will be used in the remainder of this study.

III. A Solution to the Problem

A. Clustering Procedure

We investigate the feasibility of extending the diphone database with context-sensitive diphones, in order to reduce the number of audible discontinuities. One way of keeping the database size within bounds is to cluster contexts that are spectrally alike according to a distance measure. In this part of the study, the SKL distance is used for this purpose. Suppose we divide the diphone sets C_lV $(l = 1, \cdots, M)$ and VC_r $(r = 1, \cdots, M)$, for a particular vowel V into two sets of N clusters $\{L(V)_1, \cdots, L(V)_N\}$ and $\{R(V)_1, \cdots, R(V)_N\}$, such that the maximum SKL distance across diphone boundaries in corresponding clusters $L(V)_k$ and $R(V)_k$ $(k = 1, \cdots, N)$ is below a threshold β. The maximum distance between noncorresponding clusters $L(V)_k$ and $R(V)_l$ $(k \neq l)$ will then not be limited to β. We now construct additional clusters $R(V)_{l,k}$ $(k \neq l)$, which contain the diphones of $R(V)_l$, but which are recorded with a left-side context consisting of a representative diphone in $L(V)_k$, e.g., the diphone closest to the centroid of $L(V)_k$. Instead of concatenating a diphone from $L(V)_k$ with one from $R(V)_l$, a diphone from $R(V)_{l,k}$ will be used, which will reduce the maximum SKL distance across diphone boundaries, which is hopefully lower than β, although a guarantee cannot be given in advance. This procedure will increase the number of VC diphones for a particular vowel by a factor $N(\beta)$, which is equal to the number of clusters. The database could also have been extended with left diphone clusters instead of right, but if both left and right diphones were added, the selection process would be more complicated, requiring a Viterbi-like search. For a feasibility study, this choice is not so relevant.

The number of clusters $N(\beta)$ can under certain assumptions (when the number of clusters and diphones are large enough) statistically be related to the quality improvement. If the maximum SKL distance between corresponding clusters is β, then the total number of transitions between all corresponding clusters $(\sum_{i=1}^{N(\beta)} m_i^2)$ is less or equal than the total number of diphone combinations with maximum KL distance β. This is expressed in the right-hand inequality of (10)

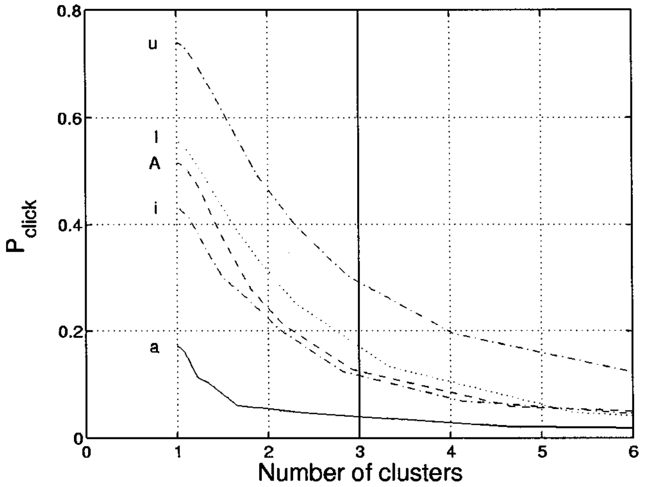

Fig. 7. Number of clusters versus P_{click}. The lines represent the lower bound, i.e., the maximum improvement that might be obtained by using additional diphones. The intersections with the vertical line indicated the maximum obtainable improvement for each vowel when using three clusters.

$$\frac{M^2}{N(\beta)} \leq \sum_{i=1}^{N(\beta)} m_i^2 \leq M^2 \int_0^\beta p(D)\, dD \tag{10}$$

with $\int_0^\beta p(D)\, dD$ being the probability of a distance smaller than β, m_i^2 the number of transitions in cluster i and M^2 the total number of transitions. It can furthermore be shown that the total number of transitions between corresponding clusters is always larger than $M^2/N(\beta)$, which is expressed in the left-hand inequality of (10). In fact, the minimum is attained when all clusters have equal sizes.

This leads to the following inequality for $N(\beta)$

$$N(\beta) \geq \frac{1}{\displaystyle\int_0^\beta p(D)\, dD\, dD}. \tag{11}$$

This factor can be used as a measure of cost of improvement. After the extension of the diphone database, the probability of an audible discontinuity occurring, P_{click}, can be computed. It constitutes two independent probabilities, one representing the probability of a discontinuity occurring in a cluster that is not detected (given by $P\{1\}[1-P_D(\beta)]$), and the other representing the probability of a detected discontinuity occurring between the original left diphone and the newly added right diphone, which is estimated to be maximally $P_D(\beta)(1-P_D(\beta))$, assuming that the distance between the new cluster and the original one is less than β. Before actual extension of the database, we can compute P_{click} by

$$P_{\text{click}}(\beta) = P\{1\}[1 - P_D^2(\beta)]. \tag{12}$$

Fig. 7 plots the number of clusters $N(\beta)$ against the probability of a click P_{click}. This represents an estimate of the maximum improvement that can be obtained by adding a certain number of clusters. The threshold β can now be chosen according to cost or performance constraints. Fig. 7 shows that when using three clusters, the probability of an audible discontinuity occurring is predicted to maximally decrease from 0.17

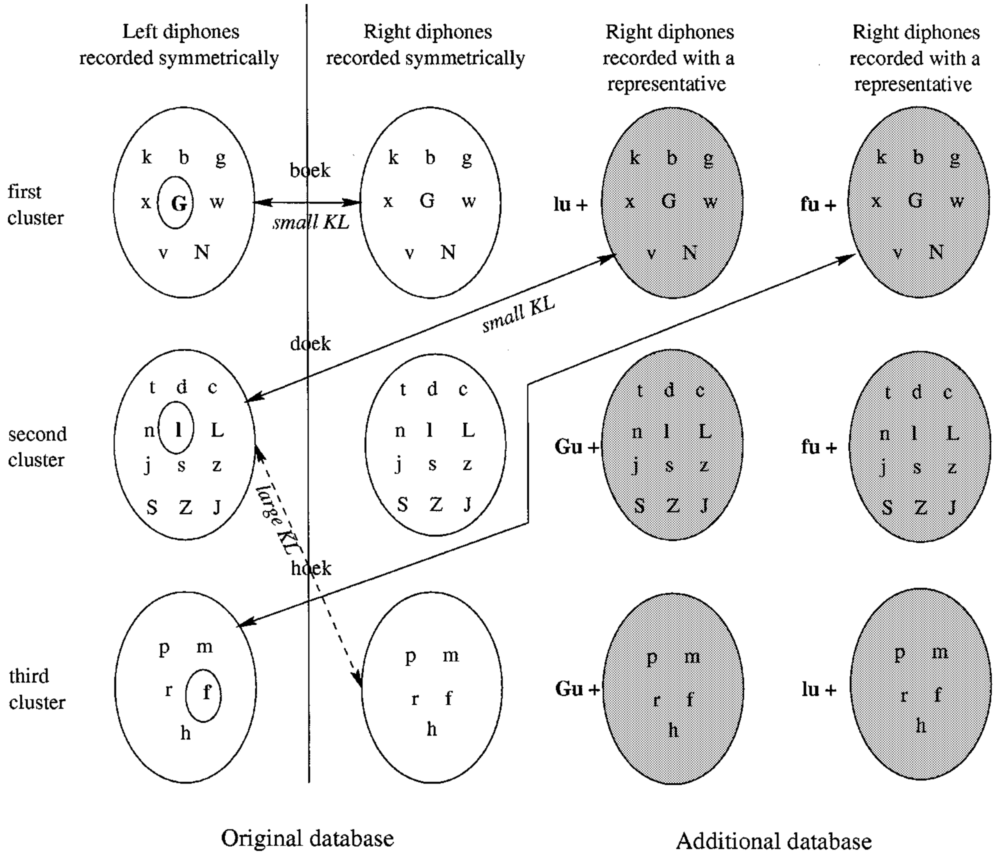

Fig. 8. The principle of the construction of additional diphone clusters. The context-sensitive diphone clusters are indicated in grey. They consist of VC$_r$ diphones that were recorded with a representative from a noncorresponding left cluster. The representative of each cluster is circled.

to 0.04 for /a:/, from 0.43 to 0.12 for /i/, from 0.52 to 0.13 for /A/, from 0.55 to 0.17 for /I/ and from 0.74 to 0.29 for /u/.

Fig. 8 illustrates the clustering procedure for the vowel /u/. In our investigation, the maximum number of clusters is restricted to three, which contain the same consonantal contexts for left and right diphones. Adding more than three clusters will not significantly reduce the probability of an audible discontinuity. Concatenating /bu/ and /uk/ to make *boek* is expected to be unproblematic, as both consonants come from the same cluster with a small SKL distance. However, for the words *doek* and *hoek* an audible discontinuity is likely to occur as they come from noncorresponding clusters. In order to remedy this, additional right diphones are recorded with a representative from a noncorresponding left cluster. In recording this means that for the /uk/ diphone which was originally recorded in the symmetrical nonsense word *k@kuk@*, two additional diphones are recorded in the asymmetrical nonsense words *l@luk@* and *f@fuk@*. Then, the word *doek* can be created by concatenating the original left diphone /du/ with the new diphone /uk/ taken from *l@luk@* and the word *hoek* is created by concatenating the same /du/ diphone with the new right diphone /uk/ coming from *f@fuk@*.

The clusters are constructed according to a classification algorithm, derived from the Linde–Buzo–Gray (LBG) algorithm [30], which is commonly used for codebook generation for the purpose of vector quantization. The SKL distance is used as a criterion for the division. A distance matrix (DM) is constructed with C_lV diphones in the rows and VC$_r$ diphones in the columns. The clustering procedure works as follows.

TABLE III
CLUSTER CONFIGURATION FOR /a:/, /i/, AND /u/. THE REPRESENTATIVES IN EACH CLUSTER ARE THE FIRST CONSONANTS IN EACH ROW

Vowel	Consonants in cluster	Maximum KL	Average KL
/a:/	1: v bfwz	1.08	0.45
	2: S	0.00	0.00
	3: x GJLNcdghjklmnprstz	1.31	0.47
/i/	1: k GNfgnpstx	2.40	0.90
	2: b JLZdjmrvwz	2.79	0.86
	3: S chl	1.85	0.89
/u/	1: G Nbgkvwx	2.36	1.08
	2: l JLSZcdjnstz	2.02	0.80
	3: f hmpr	2.39	0.90

1) Three C_lV diphones are chosen as the initial representatives of the clusters.
2) Distance matrix is reduced to a cluster matrix with SKL distances between the three representatives and the VC$_r$-diphones. Each VC$_r$-diphone is added to the cluster to which representative it has the lowest KL distance.
3) Initial representative does not necessarily have the lowest average KL distance to all other diphones in the cluster, so for each cluster a new representative is chosen that does adhere to this criterion. Then steps 2) and 3) are repeated until the cluster configuration converges.

All possible combinations of initial representatives were tried. The best ones, i.e., the ones that lead to the lowest maximal distance in a cluster, are displayed in Table III. Weighting

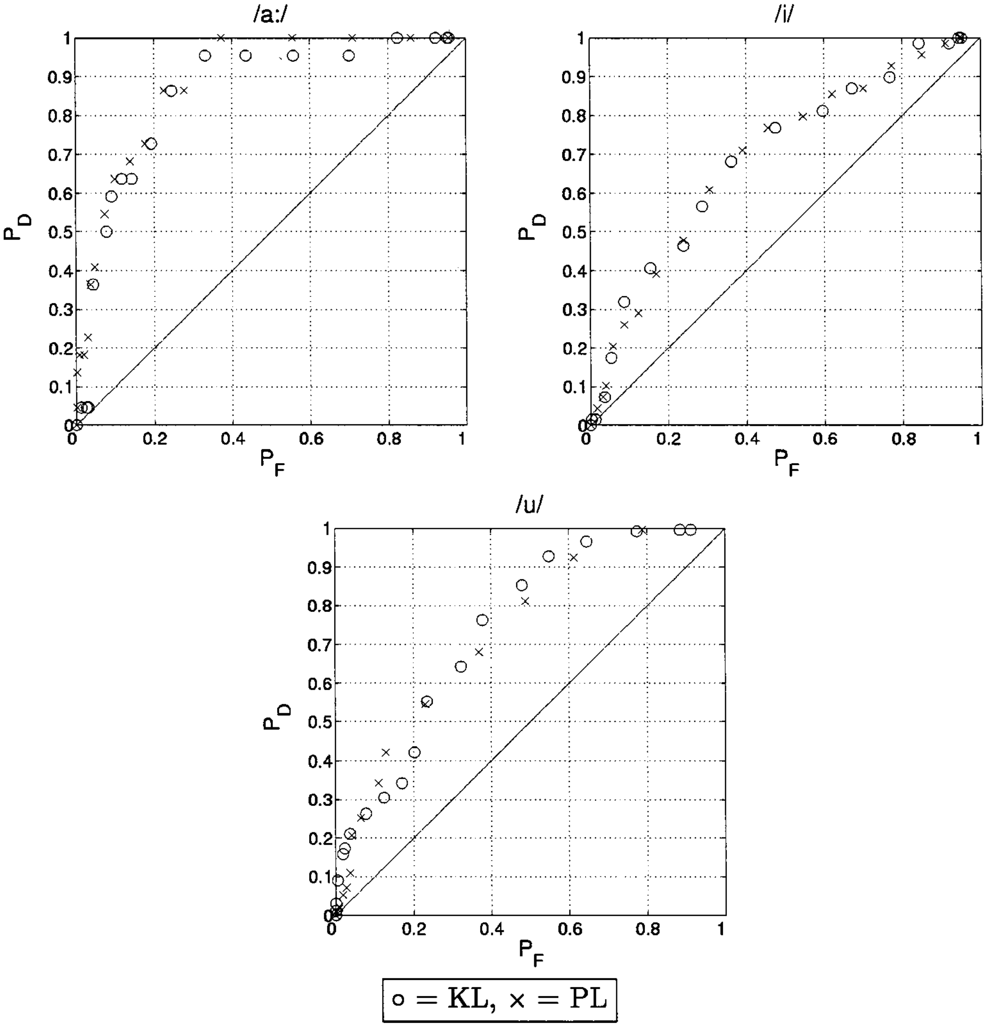

Fig. 9. ROC curves for the new database when clustering for the vowels /a:/, /i/, and /u/.

the average SKL distance with the frequency of occurrence of each diphone as measured in a large text corpus (27 000 sentences) did not have any effect on the configuration. It was decided to allow the occurrence of just one diphone in a cluster, to be able to separate an outlier that has a great spectral distance to all other diphones. For the /a:/, the /S/ ends up in a cluster on its own. There is no clear pattern related to manner or place of articulation of the consonants, except for the /u/ where all alveolars end up in the same cluster. We will come back to this issue in the discussion. The cluster configuration for /u/ was already visualized in Fig. 8.

B. Second Perceptual Experiment

In order to measure the improvement that results from the addition of context-sensitive diphones new recordings were made with which a new perceptual experiment was performed. In order to make comparison possible, the new recordings contained both the old and the new nonsense words.

1) Material: Again, stimuli were created consisting of concatenated C_lV and VC_r-diphones of the same speaker, except that now for each C_lV and VC_r combination there were two versions, one with a right diphone from the symmetrical nonsense word C_r@C_rVC_r@ (database without clustering) and one with a right diphone from the asymmetrical nonsense word C_{rep}@C_{rep}VC_r@ (database with clustering). In order to reduce the total number of stimuli, it was decided to focus on just three vowels /a:/, /i/, and /u/. The total number of stimuli used in the experiment is 2254, of which 1449 were constructed according to the original concatenation method (23 C_l × 3 V × 21 C_r) and 805 diphone combinations were obtained using diphones from the context-sensitive database (202 for /a:/, 295 for /i/ and 308 for /u/, based on three clusters).

2) Procedure: The perceptual experiment was repeated, this time using six participants with a background in psycho-acoustics or phonetics. They had not taken part in the previous experiment. The participants again had to judge whether the diphone boundary in the middle of the vowel was either smooth or discontinuous. The stimuli were presented in three hourly sessions which were held on three different days. Each session was split into two 30-min blocks by a 15-min break. The session order was different for all participants.

3) Results: Fig. 9 shows the ROC curves for the SKL and PL distances for the three vowels used in the second experiment. Table IV lists the percentage of perceived discontinuities for the new database with and without clustering. Again, these are based on the majority scores. Since we had six participants

TABLE IV
PERCENTAGE OF PERCEIVED DISCONTINUITIES PER VOWEL. THE PERCENTAGES ARE COMPUTED FROM THE SUM OF THE MAJORITY SCORES

Vowel	New database without clustering	New database with clustering
/a:/	7.7%	6.0%
/i/	19.3%	17.0%
/u/	53.4%	26.3%

TABLE V
REPEATED MEASURES ANOVA ON NEW DATABASE WITH AND WITHOUT CLUSTERING FOR SKL DISTANCE AND SUMMED PARTICIPANTS' SCORES; SIG INDICATES SIGNIFICANCE (N.S. = NOT SIGNIFICANT, * = SIGNIFICANT)

Vowel	SKL distance	Sig	Sum of participants' scores	Sig
/a:/	$F_{1,482} = 0.27$	n.s.	$F_{1,482} = 7.68$	*
/i/	$F_{1,482} = 40.02$	*	$F_{1,482} = 6.85$	*
/u/	$F_{1,482} = 65.95$	*	$F_{1,482} = 155.40$	*

in this experiment, one randomly chosen subject was left out to keep the results comparable to the old situation. The results for the new database without clustering are better than for the original database. Even though the same speaker was used as in the first experiment, there may be differences in recording conditions, speaking style or speaking rate that caused the differences. Although the clustering is based on the results of the first experiment, the results from the second experiment can be evaluated because it contains stimuli with and without clustering. Table V shows that clustering does reduce the number of audible discontinuities.

Its significance is demonstrated by a repeated measures ANOVA which was performed on the SKL distance and on the summed participants' scores. The results are presented in Table IV. When looking at the results for the KL distance one can observe that the distance has significantly decreased for context-sensitive diphones for both /i/ and /u/, but is not significant for /a:/. However, in the judgement of the participants, the number of detected discontinuities has significantly decreased for all three vowels.

When comparing the improvement prediction for the new database without clustering with the actual improvement obtained by clustering [Fig. 10], one can see that clear improvements are obtained although they are not as good as the maximally predicted improvement. The deviation from the optimal line is 0.07 for /u/ and /a:/ and 0.15 for /i/. One possible reason for this is that the maximum improvement is estimated assuming that each cluster contains an equal amount of contexts, which is not the case here.

When again considering the specific example of /duk/, it can now be seen that adding an additional /uk/ diphone that has been recorded in the appropriate context makes a noticeable difference [see Fig. 11]. Instead of an abrupt and large jump in the F_2 as observed in Fig. 1, the F_2 descends more gradually.

IV. DISCUSSION

The findings of this investigation lead to a number of interesting observations. First, the differences in results between the three vowels /a:/, /i/ and /u/ shows that /a:/ is least affected by coarticulation, whereas /i/ is more and /u/ is most affected by it. The alveolars account for most of the coarticulation in /u/. There, the slow movement of the tongue body causes the frontal mouth cavity to be small throughout the pronunciation of the vowel, which leads to a relatively high F_2 value. For /a:/ and /i/, this does not make much difference since the locus of alveolars is much nearer to their customary F_2 value than for /u/. Research by [27] and [1] has shown that coarticulation has a centralizing effect. Maybe that is the reason why /u/ is more affected in terms of F_2 and /A/ more in terms of F_1.

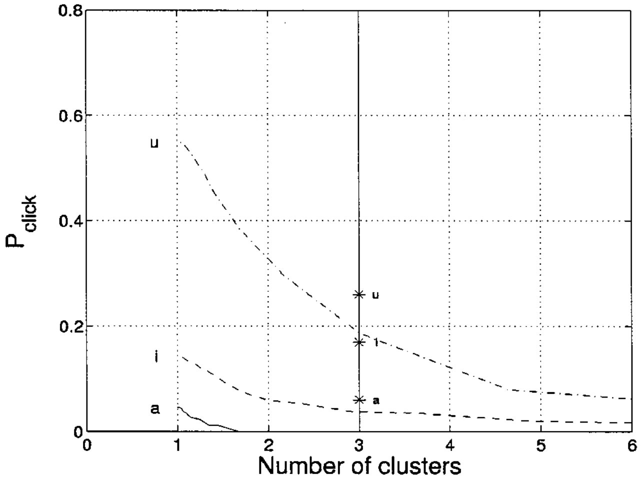

Fig. 10. Number of clusters versus P_{click}, the probability of a discontinuity arising, for /a:/, /i/, and /u/. The lines represent the lower bound, i.e., the maximum improvement that might be obtained. The stars indicated the actual improvement obtained when using three clusters.

Second, the finding that audible discontinuities still occur for the /a:/ and that clustering does not reduce the amount of audible discontinuities leads to the conclusion that besides coarticulation there is always random variation in the pronunciation of the stimuli. This was also observed by [24] who found F_2 variations in excess of 50 Hz for a vowel in repetitions of the exact same phrase as uttered by a highly professional speaker. Reference [26] reports even larger F_1 and F_2 variations (up to 250 Hz) in the repeated pronunciation of /I/ in *six* and *million* by a professional speaker. This indicates the need to record several instances of a nonsense word and choose the one that is optimal for the database.

Third, the bottom panel in Fig. 11 shows that when the diphones are recorded in asymmetrical nonsense words, the formant trajectories are no longer stable, but change gradually from start to finish. In that case, it may make sense to optimize the cutting point of the diphone boundary as proposed by [6]. Because the nonsense words used for our diphone database contain identical consonants around the vowel, it may cause the formants in the vowel to fall short of their theoretical targets. In future, it will be better to use a less constrained set of words for extracting diphones.

V. CONCLUSION

This paper reported on the occurrence of audible discontinuities in diphone synthesis caused by spectral mismatch at the diphone boundaries. A perceptual experiment was conducted to investigate the extent to which this phenomenon occurs. The results revealed that there are considerable differences between the vowels under investigation. The /a:/ showed the least amount

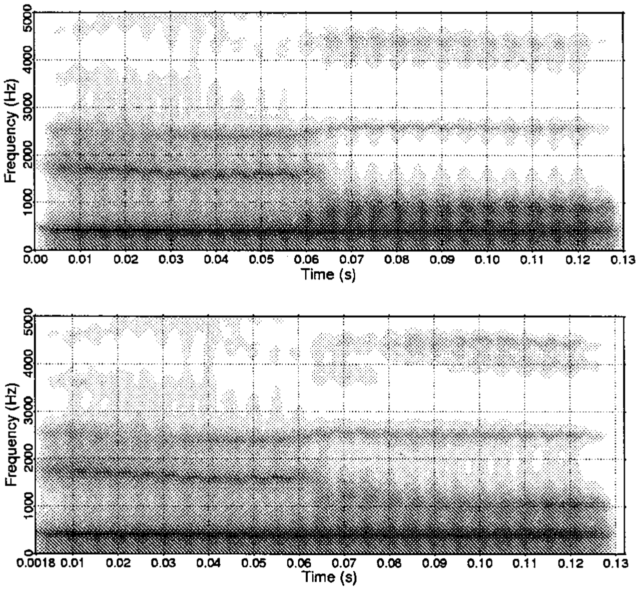

Fig. 11. Improvement in the concatenation of /du/ and /uk/ using a different diphone for /uk/. Instead of an abrupt jump in F_2 as is visible in the top panel, a gradual transition is observable in the bottom panel.

of perceived discontinuities, followed by the /i/. The /A/ and /I/ showed more perceived discontinuities, but the largest percentage of perceived discontinuities was found in the /u/. The scores obtained in the perceptual experiment were related to several spectral distance measures to find an objective measure to predict the occurrence of audible discontinuities. The relation was performed using ROC curves. The symmetrical Kullback-Leibler distance was shown to be most adequate for the task.

In the second part of this paper, a feasibility study was presented. Context-sensitive diphones were added to the database. In order to reduce the number of additional diphones, the SKL distance was used to cluster consonantal contexts that have the same spectral effects on the neighboring vowels. A second perceptual experiment was conducted to evaluate the improvement obtained with this addition to the database. A significant improvement was obtained for /u/ and /i/ both in terms of the objective SKL distance and the subjective scores. For /a:/ there was only a subjective improvement, but objectively, in terms of SKL distance, the improvement was not significant. This is not a problem, however, as the number of discontinuities in /a:/ was already low to begin with.

Although the research was performed on a restricted type of stimuli, we think the procedure of detecting audible discontinuities using the SKL measure is also applicable to other stimuli. Currently, speech synthesis using on-line selection of variable length units is very popular. We expect that the SKL measure can be successfully integrated in this approach to select the best fitting units.

References

[1] D. van Bergem, “Acoustic and lexical vowel reduction,” Ph.D. dissertation, IFOTT, Univ. Amsterdam, Amsterdam, The Netherlands, 1995.

[2] P. Boersma and D. Weenink, “Praat—A system for doing phonetics by computer,” Inst. Phonetic Sci., Univ. Amsterdam, Amsterdam, The Netherlands, http://www.praat.org, 1996.

[3] G. Booij, *The Phonology of Dutch*. Oxford, U.K.: Clarendon, 1995.

[4] P. Carvalho, L. Oliveira, I. Trancoso, and M. Viana, “Concatenative speech synthesis for European Portuguese,” in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 159–163.

[5] D. Chappell and J. Hansen, “Spectral smoothing for concatenative speech synthesis,” in *Proc. 5th Int. Conf. Spoken Language Processing (ICSLP’98)*, vol. 5, Sydney, Australia, 1998, pp. 1935–1938.

[6] A. Conkie and S. Isard, “Optimal coupling of diphones,” in *Progress in Speech Synthesis*, J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1997, pp. 293–304.

[7] R. van Dinther, P. Rao, R. Veldhuis, and A. Kohlrausch, “A measure for predicting audibility discrimination thresholds,” *IPO Annu. Progr. Rep.*, vol. 34, pp. 125–132, 1999.

[8] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*. Norwell, MA: Kluwer, 1997.

[9] E. Gigi and L. Vogten, “A mixed-excitation vocoder based on exact analysis of harmonic components,” *IPO Annu. Progr. Rep.*, vol. 32, pp. 105–110, 1997.

[10] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, 1976.

[11] H. Hermansky and J. Junqua, “Optimization of perceptually-based ASR front-end,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP’88)*, 1988, pp. 219–222.

[12] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[13] H. van den Heuvel, B. Cranen, and T. Rietveld, “Speaker variability in the coarticulation of /a, i, u/,” *Speech Commun.*, vol. 18, pp. 113–130, 1996.

[14] M. Hunt, “Signal representation,” in *Survey of the State of the Art in Human Language Technology*, R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1995, pp. 11–16.

[15] F. Itakura, “Minimum prediction residual applied to speech recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 67–72, 1975.

[16] E. Klabbers and R. Veldhuis, “On the reduction of concatenation artefacts in diphone synthesis,” in *Proc. 5th Int. Conf. Spoken Language Processing (ICSLP’98)*, vol. 5, Sydney, Australia, 1998, pp. 1983–1986.

[17] S. Krishnan and P. Rao, “A comparative study of explicit frequency and conventional signal representation for speech recognition,” *Digital Signal Process.*, vol. 6, pp. 249–284, 1996.

[18] S. Kullback and R. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[19] R. Luce and C. Krumhansl, “Measurement, scaling and psychophysics,” in *Handbook of Experimental Psychology*, S. Stevens, Ed. New York: Wiley, 1988, pp. 3–73.

[20] M. Macon, A. Cronk, and J. Wouters, “Generalization and discrimination in tree-structured unit selection,” in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 195–200.

[21] J. Makhoul and L. Cosell, “LPCW: An LPC vocoder with linear predictive spectral warping,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP’76)*, Philadelphia, PA, 1976, pp. 466–469.

[22] J. Markel and A. Gray, *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.

[23] B. Moore, B. Glasberg, and T. Bear, “A model for the prediction of thresholds, loudness and partial loudness,” *J. Audio En. Soc.*, vol. 45, no. 4, pp. 224–239, 1997.

[24] J. Olive, J. Van Santen, B. Möbius, and C. Shih, “Synthesis,” in *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, R. Sproat, Ed. Norwell, MA: Kluwer, 1998, pp. 192–228.

[25] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* Englewood Cliffs, NJ, 1993.

[26] J. Van Santen, “Prosodic modeling in text-to-speech synthesis,” in *Proc. 5th Eur. Conf. Speech Communication Technology (EUROSPEECH’97)*, Rhodes, Greece, 1997, pp. KN19–28.

[27] R. van Son, “Spectro-temporal features of vowel segments,” Ph.D. diss., IFOTT, Univ. Amsterdam, Amsterdam, The Netherlands, 1993.

[28] Y. Stylianou, T. Dutoit, and J. Schroeter, “Diphone concatenation using a harmonic plus noise model of speech,” in *Proc. 5th Eur. Conf. Speech Communication Technology (Eurospeech’97)*, vol. 2, Rhodes, Greece, 1997, pp. 613–615.

[29] J. Terken, “Spoken language interfaces: Developments,” *IPO Annu. Progr. Rep.*, vol. 31, pp. 61–65, 1996.

[30] R. Veldhuis and M. Breeuwer, *An Introduction To Source Coding*. London, U.K.: Prentice-Hall, 1993.

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

Esther Klabbers received the M.A. degree in language and computer science in 1995 from the Department of Language and Speech, University of Nijmegen, The Netherlands. In 2000, she received the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, where her dissertation was "Segmental and prosodic improvements to speech generation."

She is currently a Postdoctoral Researcher with the Spoken Language Interface Programme, IPO, Center for User-System Interaction, Eindhoven, The Netherlands. Her expertise involves many aspects of speech synthesis, such as prosodic modeling, corpus design, quality evaluation, and applications.

Raymond Veldhuis received the engineer degree in 1981 from Twente University, The Netherlands, and the Ph.D. degree in 1988 from Nijmegen University, The Netherlands, where his dissertation was "Adaptive restoration of lost samples in discrete-time signals and digital images."

From 1982 to 1998, he was a Researcher with Philips Research Laboratories, Eindhoven, The Netherlands, working in various areas of digital signal processing such as audio signal restoration and audio source coding. In 1992, he joined IPO, Center for User-System Interaction, Eindhoven. He is the author of various papers and patents in the field of sound, image, and speech processing. He is co-author of the book *An Introduction to Souce Coding* (Englewood Cliffs, NJ: Prentice-Hall) and author of the book *Restoration of Lost Samples in Digital Signals* (Englewood Cliffs, NJ: Prentice-Hall). His expertise involves digital signal processing for audio and audio source coding and speech, in particular speech synthesis. He has been active in the development of MPEG standards for audio source coding.