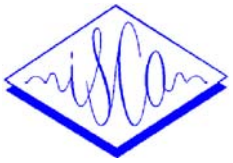

5th ISCA Speech Synthesis
Workshop
Pittsburgh, PA, USA
June 14-16, 2004

A NOVEL DISCONTINUITY METRIC FOR UNIT SELECTION TEXT-TO-SPEECH SYNTHESIS

Jerome R. Bellegarda

Speech & Language Technologies
Apple Computer, Inc.
Cupertino, California 95014

ABSTRACT

The level of quality that can be achieved by modern concatenative text-to-speech synthesis heavily depends on the optimization criteria used in the unit selection process. While effective cost functions arise naturally in the assessment of prosodic characteristics, the criteria typically selected to quantify discontinuities at the speech signal level do not tightly reflect users' perception of the resulting acoustic waveform. This paper introduces a novel discontinuity measure which jointly, albeit implicitly, accounts for both interframe incoherence and discrepancies in formant frequencies/bandwidths. This metric is derived from a distinct feature extraction paradigm, eschewing general purpose Fourier analysis in favor of a separately optimized modal decomposition for each boundary region. This alternative transform framework preserves, by construction, those properties of the waveform which are globally relevant to each concatenation considered. Experimental evaluations are conducted to characterize the behavior of the new measure, first on a contiguity prediction task, and then via a systematic listening comparison using a conventional metric as baseline. The results underscores the viability of the proposed approach in quantifying the perception of discontinuity between acoustic units.

1. INTRODUCTION

In concatenative text-to-speech (TTS) synthesis, the speech waveform corresponding to a given phoneme sequence is generated through the sequential assembly of pre-recorded speech segments. These segments, referred to as acoustic units, are normally extracted from a suitable set of sentences uttered by a professional speaker. They typically comprise variable-length phoneme or diphone sequences, whereby shorter (respectively, longer) units entail a larger (respectively, smaller) number of segment boundaries [1].

As these units are extracted from disjoint phonetic contexts, discontinuities in spectral shape as well as phase mismatches tend to occur at segment boundaries. Such artifacts usually have a deleterious effect on perception. Thus, it is generally advantageous to select longer segments in order to make synthetic speech sound more natural. Unfortunately, given the finite size of the database, the prosodic characteristics (i.e., pitch, duration, and intensity) of such candidates may not necessarily conform to the target prosodic contour [2].

The solution is to cast the segment assembly problem as a multivariate optimization task. Assuming all constraints on overall signal and prosodic behavior are suitably reflected in the associated optimization criteria, it is possible to search the available inventory of units for the optimal sequence of segments which makes up the target utterance. In practice, of course, not all constraints can be accounted for. Thus, it is typically necessary to subsequently modify the selected acoustic units in order to more precisely match the desired prosodic contour, and/or smooth the resulting signal in order to reduce audible discontinuities [3].

Many advances have recently been made in the signal processing techniques developed for such signal smoothing and prosodic modifications (see, for example, [4], [5]). This progress notwithstanding, any signal manipulation, by its very nature, is liable to degrade the acoustic waveform. It is therefore highly desirable to select units for which the minimum amount of post-processing is required [6].

This is only feasible to the extent that the optimization criteria used in unit selection exhibit a high degree of fidelity, in the sense that the metrics chosen ought to tightly predict users' perception of the resulting acoustic waveform. As far as prosodic behavior, this largely holds true. In [7], for example, each candidate unit is assessed in terms of the differences between prosodic elements such as pitch, duration, and log power, both across consecutive segments and given some target values. After allocating empirically adequate weights to each of these elements, the weighted sum proves to be a reasonable quantifier of how different units might immediately affect prosody perception.

When it comes to signal smoothing, however, things are less clear-cut. Qualitatively, the importance of various features to speech perception is well understood: for example, unnatural sounding speech typically arises from both interframe incoherence and discontinuities in the formant frequencies and in their bandwidths [8]. But quantitative measures of perceived discontinuity between two segments have proven difficult to agree upon, because they are so intricately tied to the underlying representation of speech. The latter may involve such distinct entities as FFT amplitude spectrum, perceptual spectrum, LPC coefficients, mel-frequency cepstral coefficients (MFCC), formant frequencies, or line spectral frequencies, to name but a few [9], [10]. While they are all derived from the same Fourier analysis of the signal, each entity has spawned its own metric to assess spectral-related discontinuities. In contrast, phase mismatches are typically glossed over, to be compensated for belatedly at the signal modification phase [10]. In [7], for example, the signal discontinuity between two segments involves only the cepstral distance at the point of concatenation.

This paper proposes an alternative feature extraction paradigm, which leads to a novel discontinuity metric for characterizing the acoustic (dis-)similarity between two segments. In contrast to the usual Fourier analysis, the new features are not derived via projection onto (signal-independent) complex sinusoids, but in terms of an alternative modal decomposition, separately optimized for each boundary region of interest. Because this transform framework is better suited to preserve globally relevant properties in the region of concatenation, the resulting boundary-centric represen-

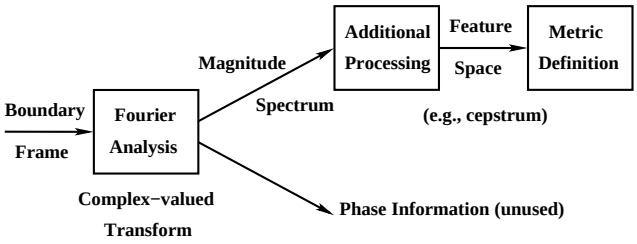

Fig. 1. *Conventional Metric Definition Framework.*

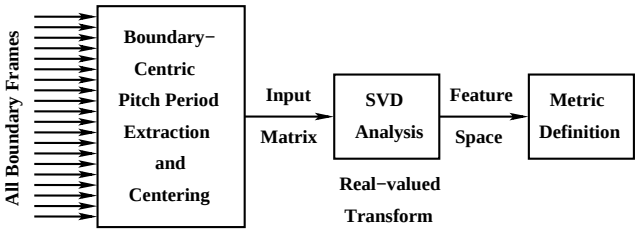

Fig. 2. *New Metric Definition Framework.*

tation proves beneficial when it comes to compare concatenation candidates against each other.

The paper is organized as follows. The next section gives a general overview of the new metric definition framework. In Section III, we describe in greater details the mechanics of the alternative modal decomposition, and discuss the different trade-offs involved. Section IV derives several discontinuity metrics which arise naturally from this representation. Finally, in Section V we report the results of two experimental evaluations, one involving a contiguity prediction task, and the other a formal listening comparison using a conventional metric as baseline.

2. OVERVIEW

Given two acoustic segments S_1 and S_2, the concatenation cost $d(S_1, S_2)$ is normally calculated from some appropriate features extracted from S_1 and S_2. From the discussion above, this is especially problematic for non-prosodic features, so this will be the focus of this paper. To successfully quantify the amount of discontinuity between S_1 and S_2, the metric $d(S_1, S_2)$ ought to tightly reflect users' perception of the resulting concatenation. Ideally, a value $d(S_1, S_2) = 0$ should 100% correlate with zero discontinuity, and a large value of $d(S_1, S_2)$ should 100% correlate with a large perceived discontinuity.

The conventional approach to this problem is depicted in Fig. 1. For each frame on either side of the boundary between S_1 and S_2, a standard Fourier analysis leads to the magnitude spectrum of the signal, while phase information is basically discarded. Optional manipulation then yields one of many spectrum-derived feature representations, such as the cepstrum. Finally, the selected representation spawns a specific spectral-related metric, such as Euclidean formant distance, symmetric Kullback-Leibler distance, partial loudness, Euclidean distance between MFCC, likelihood ratio, or mean-squared log-spectral distance, to name but a few. Many of the above spectral measures have been systematically reviewed in the literature: see, for example, [9], [10]. All tend to fall short of ideal performance: none of them succeeds in achieving a correlation with perception greater than 60-70% [10].

One possible explanation is that, when it comes to measuring perceived discontinuity, determining distances between spectral envelopes across unit boundaries may be necessary but not sufficient. Joint consideration of phase information may have to be done as well. This motivates a radical departure from the traditional Fourier analysis, involving an alternative form of "modal" analysis with simultaneous, albeit possibly implicit, treatment of both frequency and phase.

In this paper, we propose to carry out this analysis through a pitch synchronous singular value decomposition of the signal. Two observations justify this solution. First, since it is only at the boundaries that we want to measure the amount of discontinuity, all the relevant information is likely to be contained within just a few pitch periods surrounding each boundary. Hence the attractiveness of pitch synchronous processing. At the same time, when trying to decide what segment is optimal at any given boundary point, all acoustic segments straddling the boundary are likely to be germane to the decision. Hence the attractiveness of a global optimization framework such as offered by singular value analysis.

The idea is to gather, for the given boundary point, all frames in the vicinity of this point for all exemplars from the database which straddle the boundary. This leads to a matrix where each row corresponds to a particular pitch period near the given boundary. At this point, it is straightforward to perform a matrix-style modal analysis via singular value decomposition (SVD). This approach has two benefits. First, since this is a real-valued transform, both amplitude and phase information are retained, and in fact contribute simultaneously to the outcome. And second, this offers a global view of what is happening in the boundary region, as encapsulated in the vector space spanned by the resulting set of left and right singular vectors.

In fact, by analogy with the latent semantic analysis framework [11], [12], we associate with each row of the matrix (i.e., pitch period) a vector in that space. These vectors can be viewed as feature vectors analogous to, e.g., the traditional cepstral vectors. This new representation thus directly spawns new metrics $d(S_1, S_2)$ defined on the alternative feature space. The complete process is illustrated in Fig. 2.

3. MODAL DECOMPOSITION

3.1. Input Matrix

Without loss of generality, consider a (diphone-style) concatenation in the middle of the phoneme P. In the notation above, this means that the speech segment S_1 ends with the left half of P, and the speech segment S_2 starts with the right half of P. Further denote by R_1 and L_2 the segments contiguous to S_1 on the right and to S_2 on the left, respectively (i.e., R_1 comprises the second half of the P in S_1, and L_2 comprises the first half of the P in S_2). In other words, the available database is assumed to contain the segments S_1-R_1 and L_2-S_2, but not S_1-S_2.

Now let $p_K \ldots p_1$ denote the last K pitch periods of S_1, and $\bar{p}_1 \ldots \bar{p}_K$ the first K pitch periods of R_1, so that the boundary between S_1 and R_1 falls in the middle of the span $p_K \ldots p_1\ \bar{p}_1 \ldots \bar{p}_K$. Similarly, let $q_1 \ldots q_K$ be the first K pitch periods of S_2, and $\bar{q}_K \ldots \bar{q}_1$ the last K pitch periods of L_2, so that the boundary between L_2 and S_2 falls in the middle of the span $\bar{q}_K \ldots \bar{q}_1\ q_1 \ldots q_K$. (As a result, the boundary region between S_1 and S_2 can be represented by $p_K \ldots p_1\ q_1 \ldots q_K$.) For voiced speech segments, each

pitch period is obtained through conventional pitch epoch detection (see, for example, [13], [14]). For voiceless segments, the time domain signal is similarly chopped into similar, albeit constant-length, portions.

At this point, we shuffle samples around to consider, instead of $p_K \ldots p_1\ \bar{p}_1 \ldots \bar{p}_K$, the span $\pi_{-K+1} \ldots \pi_0 \ldots \pi_{K-1}$, where π_0 comprises the right half of p_1 and the left half of $\bar{p}_1$, π_{-k} comprises the right half of p_{k+1} and the left half of p_k, and π_k comprises the right half of $\bar{p}_k$ and the left half of $\bar{p}_{k+1}$ (for $1 \leq k \leq K-1$). This means that we now have $2K-1$ pitch periods instead of $2K$, with the boundary falling exactly in the middle of π_0. Similarly, we consider $\sigma_{-K+1} \ldots \sigma_0 \ldots \sigma_{K-1}$ as the centered representation for L_2-S_2, with the boundary between L_2 and S_2 falling exactly in the middle of σ_0.

Further assume that there are M segments like S_1-R_1 and L_2-S_2 present in the database, i.e., with a boundary in the middle of the phoneme P, and that for each of these we have extracted the relevant first and last K pitch periods near the boundary, and centered the outcome as described above. This results in $(2K-1)M$ pitch periods in total, encapsulating the entire boundary region. Assuming N denotes the maximum number of samples observed in each of these centered pitch periods, we symmetrically zero-pad and appropriately window all centered pitch periods to N, as necessary. The outcome is a $((2K-1)M \times N)$ matrix W with elements w_{ij}, where each row c_i corresponds to a centered pitch period, and each column t_j corresponds to a slice of time samples. This matrix W, illustrated in the left-hand side of Fig. 3, is the input matrix sought. Typically, M and N are on the order of a few hundreds, and a reasonable value for K is $K = 3$.

3.2. Feature Vectors

At this point we perform the SVD of W [15] as:

$$W = U\,S\,V^{\,T}\,, \tag{1}$$

where U is the $((2K-1)M \times R)$ left singular matrix with row vectors u_i $(1 \leq i \leq (2K-1)M)$, S is the $(R \times R)$ diagonal matrix of singular values $s_1 \geq s_2 \geq \ldots \geq s_R > 0$, V is the $(N \times R)$ right singular matrix with row vectors v_j $(1 \leq j \leq N)$, $R \ll \min(N, (2K-1)M)$ is the order of the decomposition, and $^{\,T}$ denotes matrix transposition. As is well known, both left and right singular matrices U and V are column-orthonormal, i.e., $U^{\,T}U = V^{\,T}V = I_R$ (the identity matrix of order R). Thus, the column vectors of U and V each define an orthornormal basis for the space of dimension R spanned by the (R-dimensional) u_i's and v_j's. This forms the alternative feature space sought. In essence, the rank-R decomposition (1) defines a mapping between the set of centered pitch periods and (after apropriate scaling by the singular values) the set of R-dimensional feature vectors $\bar{u}_i = u_i S$.

In contrast to more conventional approaches, the feature extraction mechanism illustrated in Fig. 3 takes a global view of what is happening in the boundary region for the phoneme P. Indeed, the relative positions of the feature vectors is determined by the overall characteristics observed in the relevant pitch periods, as opposed to an analysis restricted to a particular instance, be it frequency domain processing or otherwise. Hence, two vectors $\bar{u}_k$ and $\bar{u}_\ell$ "close" (in some suitable metric) to one another in the new feature space can be expected to reflect a high degree of similarity, and thus potentially a small amount of perceived discontinuity.

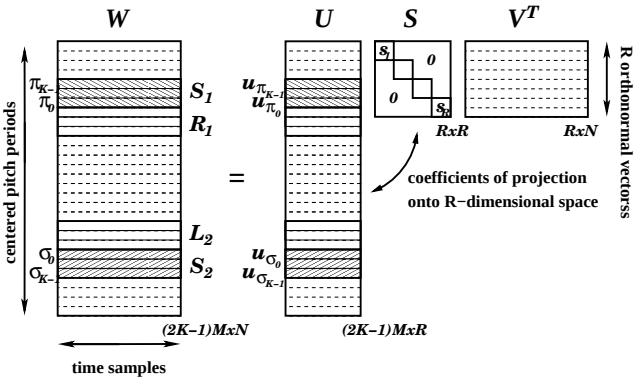

Fig. 3. *Decomposition of the Input Matrix.*

3.3. Interpretation and Trade-Offs

The above approach has interesting parallels with standard Fourier analysis. For each row $c_i = w_{i1} \ldots w_{iN}$ of W, the latter would entail:

$$X_{i\ell} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} w_{ik}\, e^{\frac{-j2\pi k\ell}{N}} \qquad \ell = 0, 1, \ldots, N-1\,, \tag{2}$$

as well as the inverse relationship:

$$w_{ik} = \frac{1}{\sqrt{N}} \sum_{\ell=0}^{N-1} X_{i\ell}\, e^{\frac{j2\pi k\ell}{N}} \qquad k = 0, 1, \ldots, N-1\,, \tag{3}$$

where $X_i = X_{i1} \ldots X_{iN}$ is the (normalized) Fourier transform vector associated with the row c_i. If we define the symmetric complex matrix Φ such that $\Phi_{k\ell} = (1/\sqrt{N}) \exp(-j2\pi k\ell/N)$, we obtain the equivalent matrix forms:

$$X_i = c_i\,\Phi\,, \qquad c_i = X_i\,\Phi^{\,H}\,, \tag{4}$$

where $^{\,H}$ denotes Hermitian transposition. In other words, Φ is (column-)orthornormal just like U and V.

The analysis (4) corresponds to the classical decomposition of the signal as a superposition of its sinusoidal projections (see, e.g., [16]). The inner product of c_i with the kth basis sinusoid in (2) has a simple interpretation as a measure of the amplitude and phase of the complex sinusoid present in c_i at the corresponding frequency. Equivalently, each component of X_i in (3) can be seen as the (complex valued) coefficient of projection of c_i onto a particular basis sinusoid.

The resulting sinusoidal transform kernel, represented by Φ, is reasonably well justified from a psycho-acoustic point of view, since the human ear acts as a kind of Fourier spectrum analyzer. On the other hand, the ear most likely is a non-linear system, whose true "analysis" parameters are yet unknown [17]. In this respect, (4) can be regarded as an approximate (linear) analysis of the acoustic signal.

Following this line of reasoning to its logical conclusion, it becomes clear that (1) simply corresponds to an alternative linear approximation, brought about by another choice of transform kernel. From (1), each row c_i of W can be expressed as:

$$c_i = u_i\,S\,V^{\,T} = \bar{u}_i\,V^{\,T}\,, \tag{5}$$

which can be interpreted as the inner product of $\bar{u}_i$ with the set of right singular vectors V. Thus, each element of $\bar{u}_i$ can be viewed

as the (real-valued) coefficient of projection of c_i onto a particular basis right singular vector. Furthermore, since, after post-multiplying by V:

$$\bar{u}_i = u_i\, S = c_i\, V\,, \tag{6}$$

the inner product of c_i with the kth right singular vector can be interpreted as a measure of the strength of the signal at the mode represented by this right singular vector. In other words, the SVD (1) embodies an alternative modal decomposition with a transform kernel represented by V.

We readily acknowledge that this alternative decomposition is most likely inferior to the Fourier decomposition as a general-purpose signal analysis tool, if only because it does not explicitly expose the concept of frequency. On the other hand, it displays several properties which seem to be attractive for the present application: (i) it is real-valued, and therefore does not require separate treatment for magnitude and phase; (ii) it is localized in time but global in scope, since it takes into account all the observations available in the region considered; and (iii) the projection basis is inherently tailored to the situation considered. In other words, it leads to an optimized (in the L_2 sense) boundary-centric representation of the problem.

4. DISCONTINUITY METRICS

4.1. Concatenation Point

To meaningfully compare two vectors $\bar{u}_k$ and $\bar{u}_\ell$ in the new SVD-derived feature space, we draw an analogy with the latent semantic analysis framework. From [12], we infer that the cosine of the angle between the two vectors is a natural metric to consider. This results in the closeness measure:

$$K(\bar{u}_k, \bar{u}_\ell) = \cos(u_k S, u_\ell S) = \frac{u_k\, S^2\, u_\ell^T}{\|u_k S\|\, \|u_\ell S\|}\,, \tag{7}$$

for any $1 \leq k, \ell \leq (2K-1)M$. This measure in turn leads to a variety of distance metrics in the feature space. But first, we have to express the concatenation point (or, more precisely, the pitch period straddling the concatenation) in this space.

Note that the feature space comprises, in particular, the vectors $\bar{u}_{\pi_k}$ and $\bar{u}_{\sigma_k}$, representing the centered pitch periods π_k and σ_k, respectively (for $-K+1 \leq k \leq K-1$). Consider now the potential concatenation S_1-S_2 of these two units, obtained as $\pi_{-K+1} \ldots \pi_1\ \delta_0\ \sigma_1 \ldots \sigma_{K-1}$, where δ_0 represents the concatenated centered period (i.e., consisting of the left half of π_0 and the right half of σ_0). This sequence will have a corresponding representation in the global vector space given by:

$$\bar{u}_{\pi_{-K+1}}\ \ldots\ \bar{u}_{\pi_1}\ \bar{u}_{\delta_0}\ \bar{u}_{\sigma_1}\ \ldots\ \bar{u}_{\sigma_{K-1}}\,. \tag{8}$$

The only vector not directly associated with a row in the original input matrix W is $\bar{u}_{\delta_0}$. However, it can easily be calculated by treating δ_0 (a row vector of dimension N) as an additional row of the matrix W. Extending the representation (5) to that additional row implies:

$$\delta_0 = u_{\delta_0}\, S\, V^T = \bar{u}_{\delta_0}\, V^T\,, \tag{9}$$

where the R-dimensional vector u_{δ_0} acts as an additional row of the matrix U. Hence the *concatenation vector*:

$$\bar{u}_{\delta_0} = u_{\delta_0}\, S = \delta_0\, V\,, \tag{10}$$

corresponds to the representation of δ_0 in the feature space.

4.2. Derived Distances

Given $\bar{u}_{\delta_0}$, the discontinuity brought about by this concatenation can be expressed as the cumulative difference in closeness before and after concatenation. Introducing the shorthand notation:

$$\tilde{K}(u_{\sigma_{-1}}, u_{\sigma_0}, u_{\sigma_1}) = \frac{K(\bar{u}_{\sigma_{-1}}, \bar{u}_{\sigma_0}) + K(\bar{u}_{\sigma_0}, \bar{u}_{\sigma_1})}{2}\,, \tag{11}$$

for the average closeness across the boundary σ_0, we can write:

$$\begin{aligned} d(S_1, S_2) = {} & 2\,\tilde{K}(u_{\pi_1}, u_{\delta_0}, u_{\sigma_1}) \\ & - \tilde{K}(u_{\pi_1}, u_{\pi_0}, u_{\pi_{-1}}) \;-\; \tilde{K}(u_{\sigma_{-1}}, u_{\sigma_0}, u_{\sigma_1})\,, \end{aligned} \tag{12}$$

which can be thought of as the relative change in similarity that occurs during concatenation. An important special case is when the two segments considered are in fact contiguous in the database, i.e., the σ's are identically equal to the π's. In this situation, it can be easily verified that, in particular, $\delta_0 = \sigma_0 = \pi_0$. We conclude that this metric exhibits the property: $d(S_1, S_2) \geq 0$, with equality if and only if $S_1 = S_2$. In other words, it is guaranteed to be zero anywhere there is no artificial concatenation, and strictly positive at an artificial concatenation point. This ensures that contiguously spoken pitch periods always resemble each other more than the two pitch periods spanning a concatenation point.

Note that this expression can be trivially generalized to encompass more than one pitch period on either side, leading to the more general expression:

$$\begin{aligned} d(S_1, S_2) = {} & \sum_{k=1}^{K-1} 2\,\tilde{K}(u_{\pi_k}, u_{\delta_0}, u_{\sigma_k}) \\ & - \tilde{K}(u_{\pi_k}, u_{\pi_0}, u_{\pi_{-k}}) \;-\; \tilde{K}(u_{\sigma_{-k}}, u_{\sigma_0}, u_{\sigma_k})\,. \end{aligned} \tag{13}$$

Because it takes into account all the pitch periods deeemed to be relevant to the boundary region, (13) can be expected to be more accurate than (12). This, in essence, corresponds to the entire trajectory difference before and after concatenation, as expressed in the SVD-derived feature space.

5. EXPERIMENTAL RESULTS

5.1. Database

All experiments were conducted using a database currently deployed in MacinTalk, Apple's TTS offering on MacOS X. This database, referred to as the Victoria corpus, was described in details in [18]. The particular portion of the corpus selected for the experiments was the so-called "Prosodic Context" sub-corpus [18].

5.2. Preliminary Experiments

To serve as proof-of-concept, preliminary experiments focused on the phoneme $P = [A]$ (in SAMPA computer readable phonetic notation, cf. [19]). Specifically, we extracted from the database all $M = 282$ instances of speech segments (in this case, diphones) with a left or right boundary falling in the middle of the phoneme $[A]$. Then, for each instance, we extracted $K = 3$ pitch periods on the left and $K = 3$ pitch periods on the right of the boundary. This led to $2K - 1 = 5$ centered pitch periods for each boundary instance, with the boundary itself being embodied by the middle row (δ_0 in the terminology of the previous section, except for the

contiguous segments where this middle row represented the likes of π_0 and σ_0). The maximum number of samples observed in these pitch periods was $N = 125$. This led to a (1410×125) input matrix comprising data relevant to the boundary region of $[A]$. We then computed the SVD of this matrix using the single vector Lanczos method (cf. [12]), with dimension set to 10, and obtained the associated feature vectors as described in Section 3.

As an initial proxy to assess the correlation between distance measure and perceived discontinuity, one possibility is to calculate how accurately the measure can predict contiguity. In other words, given a segment left of the boundary, what is the probability that the metric correctly identifies, among all possible candidates, the segment right of the boundary which is marked as contiguous in the database? This is a rather severe test, because in practice only "near-contiguity" is required. However, the ability to predict near-contiguity is obviously related to the ability to predict contiguity, and the latter test has the merit to be objective as well as simple to implement.

As discussed earlier, by construction, the distances (12) and (13) achieve perfect contiguity prediction.[1] The interest of this initial series of experiments therefore lies in how close more conventional metrics come to predict contiguity. From the literature, we know that standard spectral measures achieve a maximum correlation with perception of 60-70% [10]. Here, of course, the criterion is much stricter, since, again, it is undoubtedly sufficient to predict "near-contiguity" from a perception point of view. Nevertheless, it would be reasonable to expect that conventional measures might be able to predict contiguity a substantial fraction of the time.

As the baseline spectral distance measure, we selected a metric commonly used both to select optimal units and to segment diphones at their optimal cutting point: the Euclidean difference between MFCC vectors (see [20], [21], among others). In the experiments reported in [10], this metric performed about average as a predictor of audible discontinuity for this phone. To conform to the state-of-the-art, 39-dimensional MFCC vectors were extracted, including the usual dynamic (delta and delta-delta) features. On the data considered, this metric correctly predicted contiguity in 14.8% of all instances. This surprisingly poor result underscores the inherent limitations of spectral-only measures when it comes to quantify discontinuity. By the same token, it bodes well for the approach adopted in this paper.

5.3. Formal Listening Tests

Having established the basic validity of the method, we then performed more formal listening tests. The stimuli consisted of eight different concatenated segments S_1-S_2 consisting of three phonemes each, with a concatenation in the middle phoneme. In SAMPA notation like before, the stimuli were chosen to be: [mAn] and [sun], as examples of a concatenation in the middle of a steady spectrum vowel; [Anu], [umA], as examples of a concatenation in the middle of a steady spectrum consonant; [lOIn] and [maUs], as examples for varying spectrum vowels; and [Alu] and [Aru], as examples for varying spectrum consonants. In each case, the concatenation was done "as is" without any post-processing whatsoever, and all possible candidates available were retained, including the original contiguous segments occuring in the database. However, the latter only served as reference material, in order to ensure that a discontinuity was present in all working cases.

[1] In practice, due to rounding and other numerical errors, the measures (12) and (13) predicted contiguity in 99.7% of all instances.

The next step was to ensure that this discontinuity potentially affected only formant frequencies/bandwidths and interframe coherence. To do so, we energy-normalized all the utterances, and constrained pitch so that it was approximatively the same on both sides of every concatenation. As it turns out, the average pitch value over the data considered was approximately 200 Hz. So we imposed a simple pitch contour on each utterance, linearly decreasing from 205Hz at the beginning to 195Hz at the end, going through exactly 200Hz at the concatenation.[2] This guaranteed that any perception of discontinuity would not be due to pitch differences at the boundary. Because of the relatively small modifications involved, we used a simple pitch modification algorithm similar to TD-PSOLA [22].

For each utterance so obtained, we proceeded to characterize the boundary region exactly as above. For the baseline, we computed standard 39-dimensional MFCC vectors, including dynamic features as previously. For the boundary-centric SVD approach, we gathered the matrix W of centered pitch periods, using the same parameters as before. From the set of all utterances associated with each particular stimulus, we then set aside the best and worst artificial concatenations, as measured by: (i) for the baseline, the Euclidean distance between the two MFCC vectors straddling the boundary, and (ii) for the SVD approach, the distance (13) calculated using the relevant vectors in the feature space. Notably, there was no overlap in the set of candidates selected by the two algorithms.

The best and the worst concatenations identified for each stimulus and for each distance measure served as material for the perceptual experiments. Seven participants were selected, five generally conversant in speech processing, and two with a more advanced background in psycho-acoustics or phonetics. The experiment was divided into two hourly sessions which were held on different days, with short breaks interspersed within each session. The experiment started with a familiarization phase in which different stimuli were used to demonstrate concatenations which were clearly smooth (in fact, contiguous) and concatenations which were clearly discontinuous.

For each stimulus, the participants listened sequentially to the two groups comprising the two best and the two worst concatenations, as identified by each of the two measures. In each case, the order of presentation was randomized, both within and across groups. The subjects had to judge whether the transition at the diphone boundary was decisively smoother, about the same, or decisively more discontinuous in the first utterance than in the second. Because subjects had to concentrate on just one discontinuity, and had minimal distractions from syntactic and semantic constructs, this setup was thought to result in a more critical test than when using real speech [10]. The comparative nature of the setup was also believed to avoid the common problem of varying thresholds among listeners.

5.4. Results

The participants all felt they had been able to make consistent decisions after the familiarization phase. The results for the best concatenations are summarized in Table I. They show that the candidates selected using the SVD approach were preferred about an order of magnitude more often than those selected by the standard

[2] Clearly, a flat contour at 200Hz would have been more expedient. The reason this particular contour was selected instead was to make the test material sound a bit more natural to the evaluators [23].

Table I. Listener Preference Results for Best Concatenations. Maximum Score Achievable is 7.

Stimulus	Best Concatenations		
	Prefer SVD	Prefer None	Prefer MFCC
[mAn]	6	1	0
[sun]	5	2	0
[lOIn]	5	2	0
[maUs]	4	2	1
[Anu]	5	2	0
[umA]	4	3	0
[Alu]	3	3	1
[Aru]	3	3	1
Average Score	4.4	2.3	0.4

MFCC-based Euclidean distance metric. We infer that the SVD-selected candidates contained a smaller amount of perceivable audible discontinuity, which in turn points to a higher agreement of the SVD distance with perceived outcome.

Analogous results for the worst concatenations are omitted for space reasons, but will be presented at the workshop. In this case, the opposite phenomenon occurred, i.e., the SVD-selected candidates were preferred approximately five times less often than the MFCC-selected candidates. This indicates that the latter were better, on the average, than they should have been. Note, however, that there was a large proportion of "Prefer None" in this case, presumably due to the difficulty of judging small perceptual differences between two "bad" utterances.

This caveat notwithstanding, the evidence is consistent with a higher correlation of the SVD distance with perception. This conclusion appears to hold true particularly well for monophthongs, diphthongs, and nasals, and to a slightly lesser extent for liquids. Overall, these results confirm the suitability of the proposed approach in quantifying discontinuity between acoustic units.

6. CONCLUSION

We have proposed a boundary-centric discontinuity measure as an alternative metric to assess, at the speech signal level, smoothness (or lack thereof) between concatenated segments. Because this new distance jointly accounts for both interframe incoherence and discrepancies in formant frequencies/bandwidths, it more tightly reflects users' perception of the resulting acoustic waveform. This in turn makes it an attractive optimization criterion for use in the unit selection process.

The new metric is derived from the modal decomposition of information gathered across the entire boundary region of interest. Compared to the standard spectral transformation using the usual Fourier basis, this alternative, SVD-based feature analysis is better suited to preserve those properties of the signal which are globally relevant to the concatenation considered. For a given boundary region, the pitch periods extracted from all possible candidate units are mapped onto an separately optimized feature space of relatively low dimension. Then, each potential combination is scored in terms of the distance measure (13).

The behavior of this measure was characterized first on a simple contiguity prediction task, and then through listening experiments involving direct comparison with the standard Euclidean distance between MFCC vectors. The proposed distance was verified to correlate better with perceived discontinuity than the MFCC distance. Future efforts will concentrate on exploiting this new paradigm in other relevant areas of unit selection, such as for example the optimal training of unit boundaries.

7. REFERENCES

[1] K. Takeda, K. Abe, and Y. Sagisaka, "On the Basic Scheme and Algorithms in Nonuniform Unit Speech Synthesis," in *Talking Machines*, G. Bailly and C. Benoit, Eds., Amsterdam, The Netherlands: North–Holland, pp. 93–105, 1992.

[2] W.N. Campbell and A. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis," in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Hirschberg, and J. Olive, Eds., New York: Springer–Verlag, pp. 279–292, 1997.

[3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next–Gen TTS System," in *Proc. 137th Meeting Acoust. Soc. Am.*, 1999.

[4] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Trans. Speech Audio Proc., Special Issue Speech Synth.*, N. Campbell, M. Macon, and J. Schroeter, Eds., Vol. 9, No. 1, pp. 21–29, January 2001.

[5] J. Wouters and M. Macon, "Control of Spectral Dynamics in Concatenative Speech Synthesis," *IEEE Trans. Speech Audio Proc., Special Issue Speech Synth.*, N. Campbell, M. Macon, and J. Schroeter, Eds., Vol. 9, No. 1, pp. 30–38, January 2001.

[6] M. Balestri, A. Pachiotti, S. Quazza, P. Salza, and S. Sandri, "Choose the Best to Modify the Least: A New Generation Concatenative Synthesis System," in *Proc. Eurospeech'97*, pp. 601–604, 1997.

[7] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using Large Speech Database," in *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 373–376, 1996.

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

[9] J. Wouters and M.W. Macon, "A Perceptual Evaluation of Distance Measures for Concatenation Speech Synthesis," in *Proc. Int. Conf. Spoken Language Proc.*, Sydney, Australia, Vol. 6, pp. 159–163, 1998.

[10] E. Klabbers and R. Veldhuis, "Reducing Audible Spectral Discontinuities," *IEEE Trans. Speech Audio Proc., Special Issue Speech Synth.*, N. Campbell, M. Macon, and J. Schroeter, Eds., Vol. 9, No. 1, pp. 39–51, January 2001.

[11] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J. Am. Soc. Inform. Science*, Vol. 41, pp. 391–407, 1990.

[12] J.R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proc. of the IEEE, Special Issue Speech Recog. Underst.*, B.-H. Juang and S. Furui, Eds., Vol. 88, No. 8, pp. 1279–1296, August 2000.

[13] D. Talkin, "Voicing Epoch Detection Determination with Dynamic Programming," *J. Acoust. Soc. Am.*, Vol. 85, No. Supplement 1, 1989.

[14] Y.M. Cheng and D. O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. 37, No. 12, pp. 1805–1815, 1989.

[15] J.K. Cullum and R.A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations – Vol. 1 Theory*, Chapter 5: Real Rectangular Matrices, Boston: Brickhauser, 1985.

[16] D.C. Champeney, *A Handbook of Fourier Theorems*, Cambridge University Press, 1987.

[17] J.O. Smith III, *Mathematics of the Discrete Fourier Transform (DFT)*, W3K Publishing, 2003, ISBN 0-9745607-0-7.

[18] J.R. Bellegarda, K.E.A. Silverman, K.A. Lenzo, and V. Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Trans. Speech Audio Proc., Special Issue Speech Synth.*, N. Campbell, M. Macon, and J. Schroeter, Eds., Vol. SAP–9, No. 1, pp. 52–66, January 2001.

[19] Speech Assessment Methods Phonetic Alphabet (SAMPA), "Standard Machine–Readable Encoding of Phonetic Notation," ESPRIT project 1541, 1987–89, cf. http://www.phon.ucl.ac.uk/home/sampa/home.htm.

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

[21] 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, pp. 159–163, 1998.

[22] E. Moulines and F. Charpentier, "Pitch–Synchronous Waveform Processing Techniques for Text–to–Speech Synthesis using Diphones," *Speech Communication*, Vol. 9, pp. 453–467, 1990.

[23] K.E.A. Silverman, personal communication, 2003.