LSM–Based Feature Extraction for Concatenative Speech Synthesis

Jerome R. Bellegarda

Speech & Language Technologies
Apple Computer, Inc., Cupertino, California 95014, USA
jerome@apple.com

Abstract

In modern concatenative synthesis, unit selection is normally cast as a multivariate optimization task, yet comprehensively encapsulating the underlying problem of perceptual audition into a rich enough mathematical framework remains a major challenge. Objective functions typically considered to quantify acoustic discontinuities, for example, do not closely reflect users' perception of the concatenated waveform. This paper considers an alternative feature extraction paradigm, which eschews general purpose Fourier analysis in favor of a modal decomposition separately optimized for each boundary region. The ensuing transform preserves, by construction, those properties of the signal which are globally relevant to each concatenation considered. This leads to a join cost strategy which jointly, albeit implicitly, accounts for both interframe incoherence and discrepancies in formant frequencies/bandwidths. Systematic listening tests underscore the viability of the proposed approach in accounting for the perception of discontinuity between acoustic units.

1. Introduction

In modern concatenative text-to-speech (TTS) synthesis, the acoustic signal is generated from pre-recorded speech units, normally extracted from a large database with varied phonetic and prosodic characteristics. The selection of the best unit sequence is cast as a multivariate optimization task, which seeks to minimize an overall cost criterion across the entire database. As this criterion takes aim at human audition, it is usually composed of: (i) a target cost (how closely candidate units in the database match the specification of the target phone sequence), and (ii) a join or concatenation cost (how smoothly neighboring units flow into one another) [1]. Because any subsequent manipulation of the concatenated waveform is liable to degrade signal quality, it is highly desirable that these cost functions accurately predict user's perception of smoothness and naturalness [2].

Target cost functions typically exhibit a sufficient degree of fidelity, in the sense that the metrics chosen tend to be reasonable quantifiers of how different units might immediately affect, e.g., prosody [3]. Join cost functions, however, have proven harder to agree upon, because of a more complex relationship to speech perception. Qualitatively, the extent to which various features affect perception is well understood: for example, unnatural sounding speech results from both interframe incoherence and discontinuities in the formant frequencies and in their bandwidths [4]. But quantitatively, any measure of perceived discontinuity is 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), multiple centroid coefficients, formant frequencies, or line spectral frequencies, to name but a few [5] – [7]. While they are all derived from the same Fourier analysis of the signal, each representation has led to its own distance metric to assess spectral-related discontinuities. In contrast, phase mismatches are usually glossed over, to be compensated for belatedly at the signal modification stage [8].

This paper considers a different feature extraction paradigm, which leads to an alternative assessment of the (dis-)similarity between two acoustic units. In contrast to traditional Fourier analysis, the new features are not derived via projection onto signal-independent complex sinusoids, but in terms of a modal decomposition which yields a separately optimized set of basis components for each boundary region of interest. Because this transform framework is better suited to preserving globally relevant properties in the region of concatenation, the resulting boundary-centric representation proves beneficial when comparing concatenation candidates against each other [9].

The paper is organized as follows. The next section gives a general overview of feature extraction for concatenative TTS synthesis, and motivates the alternative outlook adopted. Section 3 describes in greater detail the mechanics of the underlying modal decomposition. In Sections 4 and 5, we show how the resulting signal representation can be naturally leveraged for unit selection TTS and unit boundary training, respectively. Finally, Section 6 reports on formal listening comparisons using conventional feature extraction as baseline.

2. Overview

As mentioned above, acoustic discontinuities are more difficult to quantify in a perceptually consistent way than prosodic discontinuities. Accordingly, we will focus on the problem of calculating the join cost between two acoustic units.

2.1. Underlying Issues

The conventional approach to this problem is depicted in Fig. 1. For the current speech segment straddling the boundary between the two units, a standard Fourier analysis produces 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. This representation in turn leads to a specific spectral-related metric, such as Euclidean formant distance, symmetric Kullback-Leibler distance, partial loudness, Euclidean distance between MFCC, likelihood ratio, mean-squared log-spectral distance, etc. Many of the above spectral measures have been systematically reviewed in the literature: see, e.g., [5] – [7], as well as [10] and the references therein. No single spectral distance was found to be best in all studies [10]. Not coincidentally, all fall short of ideal performance: none of them succeeds in achieving a correlation with perception greater than 60-70% (cf. [6]).

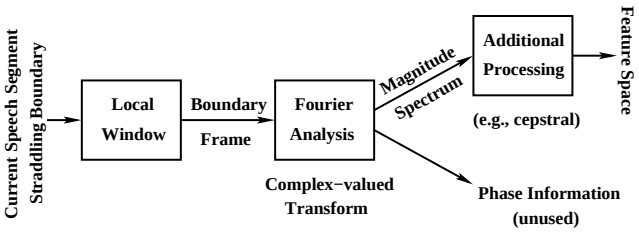

Figure 1: *Conventional Feature Extraction Framework.*

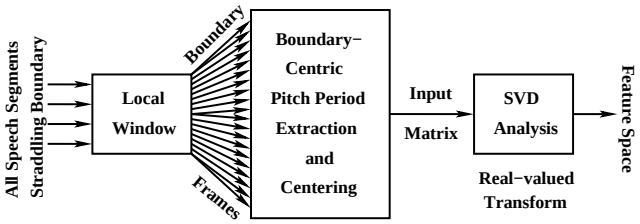

Figure 2: *Alternative Feature Extraction Framework.*

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 also be critical. This entails a radical departure from traditional (magnitude spectrum) Fourier analysis, involving an alternative form of "modal" analysis with simultaneous, albeit possibly implicit, treatment of both frequency and phase. The idea is to expose the general modes of the signal in the boundary region of interest, not just their specific frequency (or phase) components.

2.2. Alternative Framework

The implementation of this idea is guided by two observations. First, implicit or not, the treatment of phase is likely to be facilitated if we consider pitch synchronous epochs. While pitch synchronicity by itself is no panacea in the traditional Fourier framework, largely due to imperfect estimation, it is certainly worth adopting in any effort to expose general patterns in the signal. Besides, it is only at the boundaries that we want to measure the amount of discontinuity, so all the relevant information is likely to be contained within just a few pitch periods surrounding each boundary. The second observation has to do with the global scope of the analysis. When trying to decide which candidate unit is optimal at any given boundary point, all speech units straddling the boundary are likely to be germane to the decision. Thus, modes should be exposed based on features extracted, not from an individual instance, but from the entire boundary region. This assumes a global optimization framework such as offered, for example, by singular value analysis.

In this paper, we therefore adopt the framework illustrated in Fig. 2, in which the modal analysis of the signal is carried out through a pitch synchronous singular value decomposition in each boundary region of interest. For a given boundary point, we gather all frames in the vicinity of this point for all instances 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 eigenanalysis via singular value decomposition (SVD).

The idea of matrix eigenanalysis is closely aligned with the concept of non-negative matrix factorization, recently exploited in [11] for the purpose of automatic auditory scene analysis. Clearly, both kinds of decomposition aim at dimensionality reduction, for the specific purpose of exposing useful aspects of the signal. Where they differ substantially is in the input data: the matrix factorization of [11] operates on a conventional magnitude spectrogram, whereas the framework of Fig. 2 operates directly on the time-domain samples. Thus, while in [11] the rows of the matrix are associated with standard frequency bins, here they are associated with actual (untransformed) instances from the database. In other words, the SVD in Fig. 2 has the additional mission of mapping the signal into a suitable transform domain.

This outlook has several benefits. First, since the SVD is a real-valued transform, both amplitude and phase information are retained, and in fact contribute simultaneously to the outcome. 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. Third, these vectors are, by construction, optimized for this boundary region, by opposition to the traditional set of signal-independent complex sinusoids. And finally, this representation is parsimonious, to the extent that an empirically consistent value is selected for the dimension of the space. In fact, by analogy with the latent semantic analysis framework (cf. [12]), we associate with each row of the matrix (i.e., pitch period) a coordinate vector in that space, which can be viewed as a feature vector analogous to, e.g., a traditional cepstral vector. This new representation then directly leads to a concatenation metric defined on the alternative feature space.

3. Modal Decomposition

To fix ideas, consider among the set of recorded utterances the collection of all possible speech segments ending or starting within the phoneme P, so we can concentrate on a (diphone-style) concatenation within P. Two such acoustic segments, S_1-R_1 and L_2-S_2, are depicted in Fig. 3.

3.1. Matrix Construction

For both segments, we consider the boundary region consisting of the $2K-1$ centered[1] pitch periods $\pi_{-K+1} \ldots \pi_0 \ldots \pi_{K-1}$ (across S_1-R_1) and $\sigma_{-K+1} \ldots \sigma_0 \ldots \sigma_{K-1}$ (across L_2-S_2). In each case, the boundary falls exactly in the middle of either π_0 or σ_0. For voiced speech units, each pitch period is defined as the span between two consecutive glottal closure points, and obtained through conventional pitch epoch detection (see, e.g., [13]). For voiceless speech units, the time domain signal is similarly chopped into analogous, albeit constant-length, portions.

Further assume that there are M units like S_1-R_1 and L_2-S_2 present in the unit inventory, i.e., with a boundary within P. 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

[1] With a *centered* representation, the boundary can be precisely characterized by a single vector in the resulting feature space [9]. (In a more conventional framework, the boundary is normally inferred *a posteriori* from the position of the two vectors on either side.)

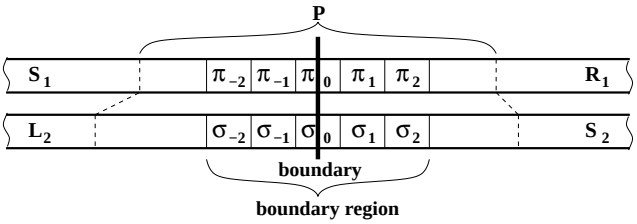

Figure 3: *Speech Segment Notation ($K = 3$).*

row r_i corresponds to a centered pitch period, and each column c_j corresponds to a slice of time samples. This matrix W, illustrated in the left-hand side of Fig. 4, is the input matrix sought.

3.2. SVD Decomposition

At this point we perform the SVD of W (cf. [9]) 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 $(1 \le i \le (2K-1)M)$, S is the $(R \times R)$ diagonal matrix of singular values $s_1 \ge s_2 \ge \ldots \ge s_R > 0$, V is the $(N \times R)$ right singular matrix with row vectors v_j $(1 \le j \le N)$, $R < \min(N, (2K-1)M)$ 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. By analogy with latent semantic analysis (cf. [12]), this space is sometimes called the *latent semantic mapping (LSM)* space $\mathcal{L}$ [14]. This is because, in essence, the rank-R decomposition (1) defines a *mapping* between the set of centered pitch periods and (after appropriate scaling by the singular values) the set of R-dimensional feature vectors $\bar{u}_i = u_i S$.

The feature extraction mechanism illustrated in Fig. 4 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 in the relevant pitch periods, and thus potentially a small amount of perceived discontinuity in the ensuing concatenated acoustic signal.

3.3. Comparison with Fourier Analysis

The above approach has interesting parallels with standard Fourier analysis. Introducing the sinusoidal transform kernel, defined as the symmetric complex matrix Φ such that $\Phi_{k\ell} = (1/\sqrt{N})\ \exp\{-j2\pi k\ell/N\}$, such analysis entails:

$$X_i = r_i\,\Phi\,, \qquad r_i = X_i\,\Phi^{\,H}\,, \tag{2}$$

where $X_i = X_{i1} \ldots X_{iN}$ is the (normalized) Fourier transform vector associated with the row $r_i = w_{i1} \ldots w_{iN}$ of W, and ^{H} denotes Hermitian transposition. Note that, in particular, Φ is (column-)orthornormal just like U and V.

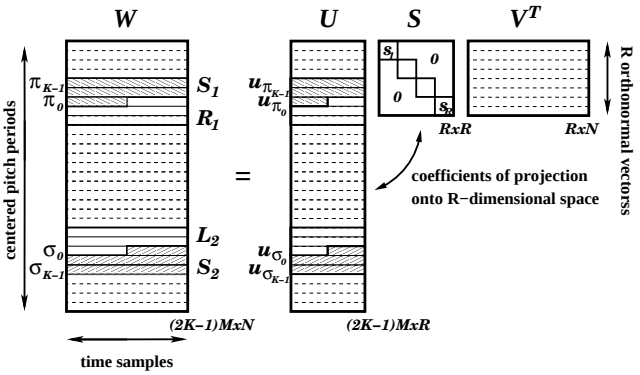

Figure 4: *Decomposition of the Input Matrix.*

The analysis (2) corresponds to the classical decomposition of the signal as a superposition of its sinusoidal projections. The inner product of r_i with the kth basis sinusoid has a simple interpretation as a measure of the amplitude and phase of the complex sinusoid present in r_i at the corresponding frequency. Equivalently, each component of X_i can be seen as the (complex valued) coefficient of projection of r_i onto a particular basis sinusoid. The sinusoidal transform kernel 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. In this respect, (2) can be regarded as an approximate (linear) analysis of the acoustic signal.

It thus becomes clear that (1) simply corresponds to an alternative linear approximation, brought about by another choice of transform kernel. From (1), each row r_i of W can be expressed as:

$$r_i = u_i\,S\,V^{\,T} = \bar{u}_i\,V^{\,T}\,, \tag{3}$$

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 r_i onto a particular basis right singular vector. Furthermore, since, after post-multiplying by V:

$$\bar{u}_i = u_i\,S = r_i\,V\,, \tag{4}$$

the inner product of r_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 transform is most likely inferior to the Fourier approach 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 unit instances which are germane to the given boundary region; (iii) the projection basis is data-driven, and hence inherently tailored to the situation considered, and (iv) the dimension of the basis vectors is inherently parsimonious (in the least squares sense). Basically, the LSM framework leads to an efficient, optimal (for the Euclidean norm), boundary-centric representation of the problem. This has potential benefits in several aspects of unit selection TTS synthesis.

4. LSM–Based Unit Selection

The first such aspect is the process of unit selection itself. Referring back to Fig. 2 and focusing on the potential concatenation S_1-S_2, we would like to make sure that this concatenation exhibits minimal discontinuities. To carry out this task, we first have to express the concatenation point (or, more precisely, the centered pitch period straddling the concatenation) in the feature space $\mathcal{L}$, and then define a suitable measure on this space.

4.1. Concatenation Point

The concatenation S_1-S_2, shown as the shaded area in Fig. 4, can be expressed 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). By construction, the feature space $\mathcal{L}$ already comprises the vectors $\bar{u}_{\pi_k}$ and $\bar{u}_{\sigma_k}$, representing the centered pitch periods π_k and σ_k, respectively (for $-K+1 \leq k \leq K-1$). This concatenated sequence therefore has a representation in $\mathcal{L}$ given by:

$$\bar{u}_{\pi_{-K+1}} \; \ldots \; \bar{u}_{\pi_1} \; \bar{u}_{\delta_0} \; \bar{u}_{\sigma_1} \; \ldots \; \bar{u}_{\sigma_{K-1}} \,, \tag{5}$$

where only one vector, $\bar{u}_{\delta_0}$, is not directly available from the LSM mapping. This vector, however, can easily be calculated by treating δ_0 (basically a row vector of dimension N) as an additional row of the original input matrix W. In fact, we trivially obtain:

$$\bar{u}_{\delta_0} = u_{\delta_0} \, S = \delta_0 \, V \,, \tag{6}$$

by simply extending the representation (4) to that additional row. Hence the *concatenation vector* (6) corresponds to the representation of δ_0 in $\mathcal{L}$.

4.2. Discontinuity Metric

Given $\bar{u}_{\delta_0}$, the discontinuity brought about by this concatenation can easily be calculated as a function of the difference in "closeness" between vectors before and after concatenation. From [12], [14], we infer that a natural measure to consider is the cosine of the angle between vectors. We therefore specify the closeness between two individual vectors as:

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

for any $1 \leq k, \ell \leq (2K-1)M$. Introducing the shorthand notation:

$$\tilde{K}(u_{\sigma_{-1}}, u_{\sigma_0}, u_{\sigma_1}) = \frac{K(\bar{u}_{\sigma_{-1}}, \bar{u}_{\sigma_0}) + K(\bar{u}_{\sigma_0}, \bar{u}_{\sigma_1})}{2} \,, \tag{8}$$

for the average closeness across the boundary σ_0, we therefore define the *discontinuity score* between S_1 and S_2 as:

$$d(S_1, S_2) = \sum_{k=1}^{K-1} 2 \, \tilde{K}(u_{\pi_k}, u_{\delta_0}, u_{\sigma_k}) - \tilde{K}(u_{\pi_k}, u_{\pi_0}, u_{\pi_{-k}}) \; - \; \tilde{K}(u_{\sigma_{-k}}, u_{\sigma_0}, u_{\sigma_k}) \,. \tag{9}$$

The discontinuity score can be thought of as the relative cumulative change in closeness that occurs within the boundary region along the entire concatenation trajectory.

An important special case is when the two speech units 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 $\delta_0 = \sigma_0 = \pi_0$. Hence $d(S_1, S_2) \geq 0$, with equality if and only if $S_1 = S_2$: 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. In other words, the closer to zero the discontinuity score, the smoother (and thus more attractive) the concatenation, and the larger the discontinuity score, the more perceptibly salient the concatenation [9].

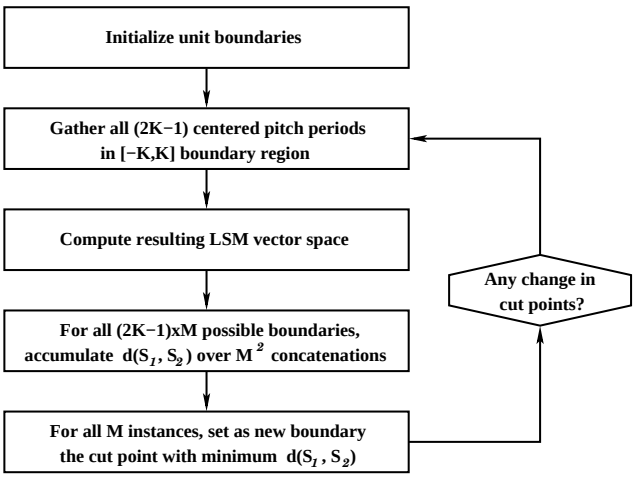

Figure 5: *Iterative Training of Unit Boundaries.*

5. LSM–Based Boundary Training

Boundary-centric feature extraction can also be leveraged for unit segmentation, at the time the underlying unit inventory is composed. This entails systematically optimizing all unit boundaries *before* unit selection, so as to effectively minimize the likelihood of a really bad concatenation. We refer to this (off-line) optimization as the data-driven "training" of the unit inventory, in contrast to the (run time) "decoding" process embedded in unit selection [15]. To carry out this task, we follow the approach recently advocated in [15], which effectively guarantees that at run time, uniformly high quality units are available to choose from.

5.1. Implementation

In [15], the evaluation criterion (9) is embedded in an iterative procedure to sequentially refine (train) the unit boundaries. The basic idea is to focus on each possible boundary region in turn, compute the LSM space associated with this region, adjust individual boundaries in that space, update the boundary region accordingly, and iterate until convergence. At each iteration, the discontinuity score (9) resulting from the concatenation of every instance of a particular unit with all other instances of that unit is computed for a neighborhood of the current hypothesized boundary. The cut point yielding the lowest average score is then retained as the new boundary for the next iteration.

The iterative boundary training procedure follows the flowchart of Fig. 5. The initialization step can be performed in a number of different ways,[2] but in practice, we have found little

[2]For example, the initial boundary for each instance can be placed in the most stable part of the phone (where the speech waveform varies the

difference in behavior based on these various forms of initial conditions [15]. Once this is done, we gather the $2K-1$ centered pitch periods for each unit instance, and derive the resulting LSM space $\mathcal{L}$. This leads to $(2K-1)M$ feature vectors in the space, and hence as many potential new boundaries. For each of them, we compute the associated average discontinuity by accumulating (9) over the set of M^2 possible concatenations. This results in $2K-1$ discontinuity scores for each instance, the minimum value of which yields the cut point to be retained. The new boundaries form the basis for a new boundary region, and the procedure iterates until no change in cut points is necessary.

5.2. Convergence

Since the boundary region shifts from one iteration to the next, the LSM space does not stay static. While this complicates the derivation of a theoretical proof of convergence, it can still be done by exploiting the fact that after each iteration the space remains relatively close to its previous incarnation. As shown in [16], the iterative procedure does converge in the least squares sense to a global minimum.

The associated final boundaries are therefore globally optimal across the entire set of observations for the phoneme P. Note that, with the choice of the LSM framework, this outcome holds given the exact same discontinuity measure later used in unit selection. Not only does this result in a better usage of the available training data, but it also ensures tightly matched conditions between training and decoding.

6. Experimental Results

We now briefly summarize some of the results we have obtained using male and female voice databases deployed in MacinTalk, Apple's TTS offering on MacOS X. Qualitatively, these databases are fairly similar to the Victoria corpus described in detail in [17]. In particular, recording conditions closely follow those mentioned in [17], though individual utterances generally differ. Complete experimental conditions, as well as additional sets of results, can be found in [9] and [16].

6.1. Unit Selection

In order to support a systematic listening comparison, we considered eight different words consisting of three phonemes each, so they could be realized from concatenated units S_1-S_2 with a concatenation in the middle phoneme. In SAMPA computer readable phonetic notation [18], the test words were chosen to be: [mAn] and [sun], as examples of a concatenation in the middle of a steady spectrum vowel; [Anu] and [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.

For each test word, stimuli were appropriately selected (using the procedure detailed in [9]) from a set of assembled utterances synthesized using a female voice database. These stimuli served as material for a perceptual experiment involving seven participants (five generally conversant in speech processing, and two with a more advanced background in psycho-acoustics or phonetics). Each evaluation session started with a familiarization phase in which reference utterances were used to demonstrate concatenations which were clearly smooth (in fact, contiguous) and concatenations which were clearly discontinuous.

Table 1. Unit Selection Listener Preference Results. Maximum Score Achievable is 7.

Test Word	Prefer LSM	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
95% Confidence	± 0.7	± 0.5	± 0.3

For each test word, the participants listened sequentially to two stimuli comprising the two concatenations identified as best by each of two measures: (i) the LSM metric (9), and (ii) the standard Euclidean distance between MFCC vectors. In each case, the order of presentation was randomized. 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 at a time, 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 [6]. The comparative nature of the setup was also believed to mitigate the common problem of varying thresholds among listeners. The participants all felt they had been able to make consistent decisions after the familiarization phase.

Tabulating the results for each test word yields the distribution of favored candidates presented in Table 1. For each column, the average score represents the average number of participants who elected the associated outcome. All confidence intervals are calculated at the 95% confidence level. Table 1 shows that the candidates selected using the LSM approach were preferred about an order of magnitude more often than those selected by the standard MFCC-based metric. Furthermore, the "Prefer LSM" outcome is significantly more likely than the combination of "Prefer MFCC" and "Prefer None" outcomes. We infer that the LSM-selected candidates contained a smaller amount of perceivable audible discontinuity, which in turn points to a higher agreement of the LSM distance with perceived outcome.

6.2. Boundary Training

A formal listening test was also performed to establish the practical validity of the iterative boundary training procedure proposed in Fig. 5. As stimuli, we generated a set of five whole sentences, where the database was segmented entirely using either of two ways. In the baseline case, unit boundaries were (classically) obtained by placing the cut point in the most stable part of the phone. In the alternative case, they were taken from the final iteration of boundary training. This resulted in two renditions of each sentence, this time synthesized using a male voice database.

Nine participants were selected, including two users with no background whatsoever in phonetics or speech processing. For

least), or, more expediently, simply at its midpoint [15].

Table 2. Boundary Training Listener Preference Results. Maximum Score Achievable is 9.

Utterance	Prefer LSM	Prefer None	Prefer Base
Example1	5	2	2
Example2	3	4	2
Example3	8	0	1
Example4	3	2	4
Example5	9	0	0
Average Score	5.6	1.6	1.8
95% Confidence	± 2.2	± 1.3	± 1.2

each pair of utterances, they were asked to listen sequentially to the two renditions, and indicate which version they preferred overall, if any. In each case, the order of presentation was randomized. After rendering each judgment, they were given a chance to verbally express what motivated their decision. Tabulating the results for each example yields the distribution of favored sentences presented in Table 2.

The five sentences were, respectively: (i) "The boy butterfly did not like the purple spot.", (ii) "Please feed the cow right away.", (iii) "It was years ago, but you still toy around.", (iv) "Investors had expected the Fed would stop before negatively affecting the economy.", and (v) "A writer who claims the manuscript copied from his work insisted in court Wednesday that there were specific echoes of his book in the best-selling thriller."

Not suprisingly, differences between the two approaches appear to be more pronounced over some segments than others. Segments most often singled out by participants included: in `Example1`, "boy butterfly" and "the purple;" in `Example2`, "the cow" and "right away;" in `Example3`, "years ago" and "toy around;" in `Example4`, "expected" and "negatively;" and in `Example5`, "a writer," "insisted in court," and "specific echoes."

Table 2 shows that, on the average, the sentences synthesized from the database featuring the optimal cut points were preferred over three times more often than those synthesized from the database with the baseline cut points. Furthermore, the "Prefer LSM" outcome is substantially more likely than the combination of "Prefer Base" and "Prefer None" outcomes. We infer that LSM feature extraction and subsequent boundary training resulted in boundaries with a smaller amount of perceivable audible discontinuity.

7. Conclusion

We have proposed a boundary-centric approach to signal representation, where the transform domain is defined via a pitch-synchronous modal decomposition of the time-domain samples gathered separately across each boundary region of interest. Compared to conventional spectral analysis using the standard Fourier basis, this alternative, LSM-based feature extraction inherently preserves those properties of the signal which are globally relevant to the concatenation considered. This makes it an attractive framework to assess smoothness (or lack thereof) between concatenated units in unit selection TTS utterances.

This boundary-centric paradigm leads to an alternative join cost strategy which jointly accounts for both interframe incoherence and discrepancies in formant frequencies/bandwidths. By leveraging both magnitude and phase information simultaneously, the resulting discontinuity metric is nominally able to reflect, more tightly than usual measures, users' perception of the concatenated acoustic waveform. This has potential benefits in several aspects of unit selection TTS synthesis, including unit selection itself, as well as the optimal training of unit boundaries.

Formal listening tests conducted in these two domains confirm that utterances synthesized using the proposed approach apparently comprise less egregious discontinuities than those synthesized in a more conventional way. This suggest that boundary-centric feature extraction is in fact well suited to quantifying perceived discontinuity between acoustic units. This conclusion seems to hold true particularly well for monophthongs, diphthongs, and nasals, and to a slightly lesser extent for liquids.

Future efforts will concentrate on more systematically exploring the influence of the feature extraction parameters (particularly K, the size of the boundary region, and R, the dimension of the feature space), in order to better characterize their relationship to factors such as phoneme identity, number of observations, dominant style of elocution, and overall prosodic context distribution.

8. References

[1] A. Hunt and A. Black, "Unit Selection in a Concatenative Speech Synthesis System Using Large Speech Database," in *Proc. ICASSP*, Atlanta, GA, pp. 373–376, 1996.

[2] M. Balestri *et al.*, "Choose the Best to Modify the Least: A New Generation Concatenative Synthesis System," in *Proc. 6th Eurospeech*, Budapest, Hungary, pp. 2291–2294, September 1999.

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

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

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

[6] 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.

[7] M. Tsuzaki and H. Kawai, "Feature Extraction for Unit Selection in Concatenative Speech Synthesis: Comparison Between AIM, LPC, and MFCC," in *Proc. ICSLP*, Denver, CO, pp. 137–140, September 2002.

[8] Y. Stylianou, "Removing Phase Mismatches in Concatenative Speech Synthesis," in *Proc. 3rd ESCA Speech Synth. Workshop*, Jenolan Caves, Australia, pp. 267–272, November 1998.

[9] J.R. Bellegarda, "A Global, Boundary–Centric Framework for Unit Selection Text–to–Speech Synthesis," *IEEE Trans. Speech Audio Proc.*, Vol. SAP–14, No. 4, July 2006.

[10] J. Vepa and S. King, "Join Cost for Unit Selection Speech Synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, S. Narayanan and A. Alwan, Eds., Upper Saddle River, NJ: Prentice Hall, pp. 35–62, 2004.

[11] P. Smaragdis, "Discovery Auditory Objects Through Non–Negativity Constraints," in *Proc. ISCA Tutorial Res. Workshop Stat. Perceptual Audio Proc.*, Jeju Island, Korea, Paper 161, October 2004.

[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] J.R. Bellegarda, "Latent Semantic Mapping," *Signal Proc. Magazine, Special Issue Speech Technol. Syst. Human–Machine Communication*, L. Deng, K. Wang, and W. Chou, Eds., Vol. 22, No. 5, September 2005.

[15] J.R. Bellegarda, "LSM–Based Boundary Training for Concatenative Speech Synthesis," in *Proc. ICASSP*, Toulouse, France, May 2006.

[16] J.R. Bellegarda, "Further Developments in LSM–Based Boundary Training for Unit Selection TTS," in *Proc. InterSpeech*, Pittsburgh, PA, September 2006.

[17] J.R. Bellegarda *et al.*, "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.

[18] 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.