Join Cost for Unit Selection Speech Synthesis

Jithendra Vepa[*], *Simon King*
Centre for Speech Technology Research
University of Edinburgh
Edinburgh, UK
www.cstr.ed.ac.uk
vepa@cstr.ed.ac.uk, Simon.King@ed.ac.uk

1 Introduction

In unit-selection speech synthesis systems, synthetic speech is produced by concatenating speech units selected from a large database or *inventory*, which contains many instances of each speech unit with varied prosodic and spectral characteristics. Hence, by selecting an appropriate sequence of units, it is possible to synthesise highly natural-sounding speech. The selection of the best unit sequence from the database is typically treated as a search problem in which the best sequence of *candidates* from the inventory is the one that has the lowest overall *cost* [11]. This cost is often decomposed into two costs: a *target cost* (how closely *candidate units* in the inventory match the specification of the *target* phone sequence) and *join cost* (how well neighbouring units can be joined) [11]. If, as is usually the case, the *cost functions* used to compute these costs only take into account properties of the fixed target sequence and local properties of the candidates, the optimal unit sequence can be found efficiently by a Viterbi search for the lowest cost path through the lattice of the target and join costs.

In this chapter we focus on the calculation of the *join cost*. The ideal join cost is one that, although based solely on measurable properties of the candidate units – such as spectral parameters, amplitude and F0 – correlates highly with human listeners' perceptions of discontinuity at concatenation points. In other words: the join cost should predict the degree of perceived discontinuity. We will use this terminology: a join cost is computed using a *join cost function*, which generally uses a *distance measure* on some *parameterisation* of the speech signal.

Usually, the join cost function contains a component which measures differences in the spectral properties of the speech either side of a proposed join between two candidate units. Other components of the function may consider differences in F0 and amplitude, for example. We concentrate here on the spectral component of the join cost function. The spectral representation is often a smoothed spectral envelope – i.e. an estimate of the vocal tract frequency response – possibly as a transformed set of coefficients, which is derived from a short-term (frame-based) analysis of the speech signal, which may optionally be pitch-synchronous. Examples of such parameterisations are: Linear Prediction (LP) coefficients; LP spectrum;

[*]Jithendra Vepa was partially funded by Rhetorical Systems Ltd.

Mel Frequency Cepstral Coefficients (MFCC); Line Spectral Frequencies (LSF); Perceptual Linear Prediction (PLP) coefficients; PLP spectrum; Multiple Centroid Analysis (MCA) coefficients. For all but MCA, see [10]; for MCA, see [6].

To measure the difference between two vectors of such parameters we need a distance measure. This may be a *metric*, provided it has the required properties[1], but this is not necessary. Examples of such measures are: absolute magnitude distance; Euclidean distance; Mahalanobis distance – i.e. Euclidean distance normalised for (co)variance; Kullback-Leibler (KL) divergence. All but the KL divergence are metrics. We have used a symmetrical version of KL divergence to compute the distance between two speech parameterisations, as explained in section 3.3.

One approach to comparing alternative join cost functions is to implement each function in a synthesiser, and compare the synthesised speech of the two systems. This is time-consuming, and requires repeated perceptual listening tests each time the join cost is altered. We use a methodology which only requires a single set of listening tests but still allows objective comparisons to be made between alternative join cost formulations. In this method, synthetic speech stimuli are generated in which there are a range of qualities of join. These stimuli are then rated by listeners. Comparison of join cost functions is then achieved by computing the correlations between join costs and listener ratings. Good join costs will correlate strongly with listener ratings of perceptual join discontinuity.

In this chapter, we first review previous work in which many combinations of *parameterisation* and *distance measure* are used as a *join cost function*. We then go on to look at alternatives, such as the use of the original context of candidates from the inventory. We then present results of our own research which uses the correlation between join cost and listener ratings to compare potential join cost functions. Finally, we touch on our latest work, which examines a single underlying representation to simultaneously compute join cost and smooth the coefficients used for waveform generation.

2 Previous Work

Since Hunt and Black [11] first described unit selection synthesis as a search for a low cost candidate unit sequence, many join cost functions have been proposed. The usual evaluation method for a proposed join cost function is some form of perceptual experiment.

2.1 Join cost functions based on spectral measures

Previous studies

Klabbers and Veldhuis [13, 14] examined various spectral distance measures for joins in five Dutch vowels to find which measure best predicts the concatenation discontinuities. These various measures were correlated with the results of a listening experiment, in which listeners have to make a choice between 0 or 1 based on whether the concatenation was smooth(0) or discontinuous(1). They found that a **symmetrical Kullback-Leibler measure on LPC power-normalised spectra** is the best predictor among their six spectral distance measures: Euclidean distances between formants (F1 and F2), MFCC, the Likelihood Ratio (LR) and the mean squared log-spectral distance (MS LSD), loudness difference (LD) and expectation differences (ED).

[1]symmetrical; triangular inequality property

A similar study by Wouters and Macon [29] showed that a **Euclidean distance on Mel-scale LPC-based cepstral parameters** is a good predictor of perceived discontinuity by evaluating several distance measures using perceptual data obtained from listening tests. In these tests, pairs of synthetic mono-syllabic English words that were made from identical unit sequences, except for one half phone, were presented to listeners to rate the difference between the word pairs on a five-point scale. The substituted half phones were limited to three specific cases of vowels. They computed the correlation between objective distance measures and mean listener responses. Their results indicated that parameterisations that use a non-linear frequency scale (such as Mel and Bark scales) performed better than those that do not. They also found that weighting individual parameters of cepstra, LSF or delta coefficients could improve correlations.

A variety of acoustic transforms (LPC, linear prediction cepstral coefficients, LSF, MFCC, residual MFCC, bispectrum, modified Mellin transform of the log spectrum, segmental modified Mellin transform of the log spectrum and Wigner-Ville distribution based cepstrum) were compared by Chen & Campbell [4] for use in assessment and evaluation of synthetic speech. The speech material was synthesised using the CHATR speech synthesis system [3]. They first segmented the original speech signal and the synthetic speech signal into frames, each frame represented by several feature coefficients. Then they used dynamic time warping (DTW) for aligning synthetic and natural segments. The overall distortion obtained from the DTW was used as a distance between the synthetic speech and natural speech. Finally, they correlated the distances computed from various acoustic transforms with listener ratings obtained from a mean opinion score (MOS) evaluation. Their results showed that the distances computed using the **bispectrum** had the highest degree of correlation with the MOS scores.

Stylianou and Syrdal [23] conducted a psychoacoustic experiment on listeners' detectability of signal discontinuities in concatenative speech synthesis. They used an experimental version of the AT & T Next-Generation system [1] to synthesise the test stimuli. In this study, the concatenative costs derived from various objective distance measures were compared with listeners' detection scores. And these distances were evaluated based on the detection rate, the Bhattacharya measure of separability of two distributions, and Receiver Operating Characteristics (ROC) curves. Their results showed that a **symmetrical Kullback-Leibler (KL) distance between FFT-based power spectra** and the **Euclidean distance between MFCC** have the highest prediction rates. In contrary to [13, 14], this study found that KL distance based on LPC spectra was the one of the worst performers.

Another study by Syrdal [24] is focused on phonetic analysis of concatenation discontinuity detection in synthesised speech using either male or female voice. She observed that these discontinuities were significantly more audible for the female voice than the male voice. However, there were many similarities between the results for the two genders. She also observed a higher rate of detection of discontinuities for diphthongs than for monophthong vowels. This is intuitively understandable, since diphthongs have a spectrally dynamic point in the middle, which is usually a join point.

Donovan [8] proposed a new join cost which can be described as a decision-tree-based context-dependent **Mahalanobis distance on perceptual cepstral coefficients**. He conducted listening tests to compare the performance of this new method with other join costs derived from Itakura and KL distances on Mel-binned power spectral; Euclidean and Mahalanobis distances on cepstra, perceptually modified MFCC (P-Cep), log energy and the first and second time differentials of cepstra and P-Cep. The test stimuli were synthesised in a male voice using a modified form of the IBM trainable speech synthesis system [7]. The correlation results showed that this new measure out-performs other measures. Also, further listening tests have justified the use of this measure in the IBM synthesis system.

Using the join cost function to enrich the inventory

Klabbers and Veldhuis studied the feasibility of extending a diphone inventory with context-sensitive diphones to reduce the occurrence of audible discontinuities [14, 15]. In order to reduce the number of additional diphones, they used their best join cost function – a **symmetrical Kullback-Leibler distance (SKL) on LPC power-normalised spectra** – to cluster the consonantal contexts that had the same spectral effects on neighbouring vowels. To evaluate the improvements gained with this extended inventory, they conducted a further perceptual experiment and observed that these additional diphones significantly reduced the number of audible discontinuities.

A method for enhancing the quality of synthetic speech by reducing the spectral mismatches between concatenated segments was recommended by Founda et al [9]. First, they used a listening test to determine which join cost function best predicted the audible discontinuities in synthetic speech. They found that a Kullback-Leibler (KL) divergence on power-normalised spectra was the best predictor among the measures they tested. They then used this distance measure to enrich their diphone database and observed a significant reduction of the spectral mismatches.

Summary

If there is a single conclusion that can be drawn from the above results, it is that no single join cost function was found to be best in all studies! It is not clear whether this is because the experimental materials vary (small sets of vowels in isolated words, for example), or because join cost is language, accent, or speaker dependent. The latter would have serious implications for synthesisers with numerous languages and voices, such as *rVoice* [2]. We look at this problem further in section 5. Another conclusion that we can draw is that the use of parameterisations that include a perceptually-motivated non-linear frequency scale is generally a good idea – a finding consistent with the types of parameterisation used in automatic speech recognition [10].

2.2 Combined join cost and target cost functions

It is well known that the target cost function determines how well a unit's phonetic contexts and prosodic characteristics match with those required in the synthetic phone sequence. Thus, usually the target cost is computed as the weighted sum of the differences between prosodic and phonetic parameters of target and candidate units. However, attaining the balance between the target cost and the join cost is not easy(i.e. if we give more emphasis on join cost, then target cost may be weighted low and thus resulting in bad synthesis). One way of lessening this behaviour is to combine these two costs. Here, we present some studies in which these two costs are combined and individual components are weighted based on perceptual/MOS experiments.

[2]TTS engine from Rhetorical systems Ltd

Functions composed of weighted sub-costs

Chu and Peng [5] presented a concatenative cost function (weighted sum of several component-costs) as an objective measure for naturalness of synthesised speech. This cost function has seven component-costs: these can be derived directly from the input text and the speech database. A formal mean opinion score (MOS) experiment was conducted in order to know the performance of this objective measure. They optimised weights for the components of this cost function using mean opinion scores and achieved high correlation (-0.872) between objective and subjective measures.

More recent work by Peng, Zhao and Chu [20] describes a method for optimising the cost function for unit selection. They used contextual information for all units used in the MOS evaluation obtained in the previous experiment to re-calculate concatenation cost with a new cost function. Then they performed optimisation in three steps. First, they optimised the cost function parameters individually. In a second step, interactions among these parameters were explored and they added higher order components to the cost function. Finally, they optimised weights for each component of the new cost function. To validate this new cost function, correlation with MOS was used. They achieved significant improvement in correlation (-0.822 to -0.897) between cost and MOS.

Minkyu Lee [17] proposed a new method for unit selection in a large corpus-based concatenative synthesis based on a perceptual preference test. His algorithm searches a set of weights for components of cost function that can produce rankings of renditions that are close to the perceptual test results. The downhill simplex method was used for this multi-dimensional search. A dissimilarity measure was employed to evaluate the closeness of two rankings, obtained from the perceptual test and his algorithm. He found that his unit selection algorithm using the optimised weights chosen the same rendition as human listener preferred in about 83 percent of the cases (5 out of 6 words).

Phonetic features

Blouin et al [2] presented a join cost function based on phonetic and prosodic features. This function is defined as a weighted sum of sub-costs, each of which is a function of various symbolic and prosodic parameters. Weights were optimised using a multiple linear regression as a function of an acoustic measure of concatenation quality. This acoustic measure is calculated as a Kullback-Leibler "distance" on normalised LPC power spectra. Perceptual evaluation results indicated that the concatenation sub-cost weights determined automatically were better than hand-tuned weights, with or without applying $F0$ and energy smoothing after unit concatenation.

A comparison of various spectral features – auditory-based (AIM) [19] features, LPCs and MFCCs – was presented by Tsuzaki & Kawai [25]. In their perceptual experiment, the listeners' task was to distinguish natural speech and synthetic speech with units selected according to the various possible join cost functions. Their results showed that the AIM-based join cost functions have a significant advantage over LPCs and similar performance to those using MFCCs.

In a parallel study, Kawai & Tsuzaki [12] compared acoustic measures and phonetic features in their ability to predict audible discontinuities. The acoustic measures were derived from MFCCs, mainly Euclidean distances between MFCCs of certain frames. A perceptual experiment was used to measure the degradation in naturalness due to signal discontinuities. Then, models were built to predict the degradation scores from the acoustic measures and phonetic features. The models used were: multiple regression model; decision tree; neural network. The multiple regression coefficients were calculated under open and closed conditions of modelling and for acoustic measures and/or phonetic features. Phonetic features were found to be more efficient than acoustic measures in predicting the audible discontinuities.

Summary

The above studies mainly address the issue of weighting individual components of the join cost function by conducting some perceptual experiments. Except a study, in which the weights were optimised based on an acoustic measure of quality. One conclusion we can draw here is that either we can use a direct acoustic measure or a measure based on phonetic and prosodic features which correlates well with human perception.

3 Spectral Distances

As we have already noted, a join cost function consists of a *distance measure* which operates on some *parameterisation* of the final and initial frames of two units to be concatenated, as shown in figure 1. A wide variety of distance measures and parameterisations are possible.

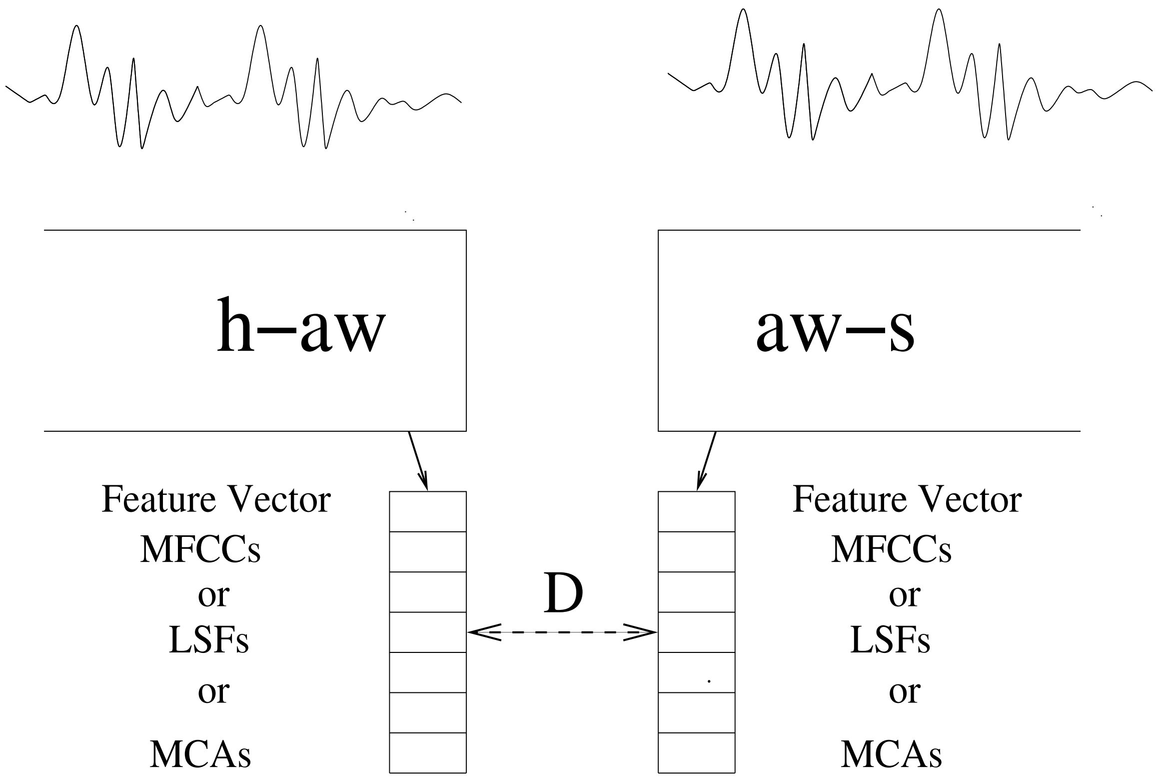

Figure 1: Objective spectral distances

Consider two diphones, *h-aw* and *aw-s*, which form a unit sequence. To compute the spectral distance between these two diphones, feature vectors of the final frame of diphone *h-aw* and the initial frame of diphone *aw-s* have to be obtained. Usually, the computation of these feature vectors is done off-line to reduce the computation cost of the system at run time. This requires storage of the feature vectors – so compact representations are preferable. Once these features are computed, various distance measures can be applied to them.

In our study, we used three parameterisations: Mel frequency cepstral coefficients (MFCCs), line spectral frequencies (LSFs) and multiple centroid analysis (MCA) coefficients. Distances are computed between pairs of these feature vectors by applying a distance measure.

3.1 Parameterisations

Mel-frequency coefficients (MFCC) [10] are a representation defined as the real cepstrum of a windowed short-time signal derived from the fast Fourier transform (FFT) of the speech signal. The difference from the real cepstrum is that a nonlinear perceptually motivated frequency scale is used, which approximates the behaviour of human auditory system. Because of their non-linear scale, decorrelated nature and robustness to noise, this representation is widely used in speech recognition.

The line spectral frequencies (LSF) [21] can be computed from an all-pole model of the speech. These occur in pairs and each pair approximately corresponds to a resonance of the vocal tract. The LSF representation has a number of properties including a bounded range, a sequential ordering of the parameters and a simple check for the filter stability which make it attractive for quantisation of LPC parameters amongst other things.

Multiple centroid analysis was introduced by Crowe & Jack [6] as an alternative to traditional formant estimation techniques, employs a global optimisation based on a generalisation of the centroid. To compute centroids, first we need to split a multi-modal distribution such as a speech power spectrum into appropriate number of partitions say 4 or 5. Then, the centroid of a specific partition of the distribution $P(n)$ bounded by $n = c_1$ and $n = c_2$ is estimated as the value that gives minimum squared error, as shown in the equation below:

$$e(c1, c2, k) = \sum_{c_1}^{c_2} (n-k)^2 P(n) \tag{1}$$

If the spectral distribution within a single partition contains a single formant then the centroid and associated variance represents the formant frequency and bandwidth [30]. This method is more robust than peak picking, so is an attractive alternative to linear prediction based formant trackers. This method also can be applied to other speech sound classes, not just those with formants.

Delta coefficients

Delta coefficients are first order time derivatives of any of the above speech parameters. These are computed in terms of the corresponding static parameters, as shown below:

$$d_t = x[t] - x[t-1] \tag{2}$$

More sophisticated methods can also be used to compute these delta coefficients.

3.2 Simple distance measures

Absolute distance

Simple absolute distance is computed as the sum of the absolute magnitude difference between individual features of the two feature vectors:

$$D_{abs}(X,Y) = \sum_{i=1}^{N} |X_i - Y_i| \tag{3}$$

Euclidean distance

The Euclidean distance between two feature vectors, X and Y is computed as shown below:

$$D_{Eu}(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{4}$$

This distance is very easy to compute. However, the Euclidean distance does not take any account of variances or covariances of the distribution of the feature vectors.

3.3 Statistically-motivated distance measures

Popular distance measures from the field of statistics include the Kullback-Leibler divergence (which is not a metric) and the Mahalanobis distance metric. We have no prior preference for any of these: we will find out which is best by experiment.

Mahalanobis distance

Mahalanobis distance, also used by Donovan in a join cost function [8], is a generalisation of standardised (Euclidean) distance in that it takes account of the variance or covariance of individual features. In general, the Mahalanobis requires the estimation of covariance matrices; often, the off-diagonal elements of the covariance matrix are assumed to be zero – this saves computation and storage.

In preliminary experiments, we found that making this diagonal covariance assumption was reasonable: using full covariance matrices did not improve performance over using diagonal matrices. The Mahalanobis distance for two feature vectors X and Y, with diagonal covariance matrix, is shown in the equation 5.

$$D_{Ma}(X,Y)^2 = \sum_{i=1}^{n} \left[\frac{X_i - Y_i}{\sigma_i}\right]^2 \tag{5}$$

where, σ_i is standard deviation of the i^{th} feature of the feature vectors, i.e. the diagonal entries of the covariance matrix.

Kullback-Leibler (KL) divergence

The Kullback-Leibler (KL) divergence [16] computes the “distance” between two probability distributions. It has been used in join cost functions before [8, 9, 13, 23], and requires us to interpret spectral envelopes as probability distributions. Let $P(\omega)$ and $Q(\omega)$ denote two spectral envelopes, then KL divergence is defined as:

$$D_{KL}(P,Q) = \frac{1}{2\pi} \int_{0}^{2\pi} P(\omega)\ log \frac{P(\omega)}{Q(\omega)} d\omega \tag{6}$$

The above divergence is asymmetric, so we can define a symmetric KL "distance" (not a metric) as:

$$\begin{aligned} D_{SKL}(P,Q) &= (D_{KL}(P,Q) + D_{KL}(Q,P))/2 \\ &= \frac{1}{4\pi} \int_{0}^{2\pi} (P(\omega) - Q(\omega))\ log \frac{P(\omega)}{Q(\omega)} d\omega \end{aligned} \tag{7}$$

The standard procedure for evaluating the above equation is by performing the integral as a summation over discrete frequencies. However, Veldhuis & Klabbers [26] recently showed that this approximation is inferior to the exact method for computing symmetrical KL distance, equation 7, for all-pole (LPC) spectra. The computational cost of this exact method is substantially higher. We used the discrete summation approximation:

$$D_{SKL}(X,Y) = \sum_{i=1}^{N} (X_i - Y_i)\ log \frac{X_i}{Y_i} \tag{8}$$

This is only valid if i is a frequency index. Hence, we have not used this distance for MFCCs. Other distance measures used in previous studies [8, 29, 23] include the Itakura-Saito and Bhattacharya distances, although neither has been shown to give good performance in join cost functions.

3.4 Weighted Distances

Each of the above combinations of parameterisation plus distance measure (i.e. join cost functions) will be measuring slightly different properties of the speech signal. It is natural to ask whether a weighted sum of several of them could perform better than any individual one. Our preliminary studies showed that a weighted sum of distance measures on MFCC, LSF and MCA can result in higher correlations with human perception than any individual measure. A weighted distance (L) of various distances can be defined as:

$$L = \sum_{j=1}^{J} w_j * D_j(X,Y) \tag{9}$$

where, w_j is weight on distance D_j between two feature vectors, X and Y.

Finding weights

The above set of equations is over-determined (i.e. more equations than unknowns), as we have more perceptual data than number of weights (i.e. three weights; each for MFCCs, LSFs and MCA coefficients). Hence, there is no exact solution for this set of equations. However, we can compute the solution as one that comes closest (e.g. in least square sense) to satisfying all equations simultaneously. To compute these weights the above equations can be re-arranged in the form of linear system of equations as shown below,

$$\mathbf{A} \cdot \mathbf{w} = \mathbf{l} \tag{10}$$

where **A** is M-by-J distance matrix and **l** is the column vector $\{L_1, \cdots, L_M\}$. We used standard least-squares method to find **w**.

4 Perceptual listening tests

The ideal join cost should correlate highly with human perception of discontinuity at concatenation points. Since the design and evaluation of a join cost function is an iterative process, it is not usually practical to conduct listening tests for each and every proposed function. It is more efficient to use a single listening test to obtain listeners' ratings of a range of synthetic stimuli, and then design the join cost function to maximise correlation with those ratings. This methodology has been used before [8, 23]; we propose a variation on the method by deliberately including a *range* of qualities of join and using *natural sentences* rather than isolated words.

This listening test measures the degree of **perceived** concatenation discontinuity in natural sentences generated by the state-of-the-art speech synthesis system, **rVoice**[3] from Rhetorical Systems Ltd. rVoice is a general purpose Text-to-speech synthesis (TTS) engine which delivers natural sounding synthetic speech. It is as good as any commercially available synthesis system.

4.1 Test stimuli

Preliminary informal listening tests indicated that spectral discontinuities are particularly prominent for joins in the middle of diphthongs, presumably because this is a point of spectral change (due to changing vocal tract shape, and therefore moving formant values). Previous studies have also shown that diphthongs have higher discontinuity detection rates than long or short vowels [24]. Our study therefore focused on joins in diphthongs.

We selected two natural sentences for each of five American English diphthongs: ey (*ei*) , ow (*oʊ*) , ay (*aɪ*), aw(*aʊ*) and oy (*ɔi*). One word in the sentence contained the diphthong in a stressed syllable. The sentences are listed in table 1.

diphthong	*sentences*
ey	More **plac**es are in the pipeline. The government sought autho**riz**ation of his citizenship.
ow	European shares resist **glob**al fallout. The speech sym**pos**ium might begin on Monday.
ay	This is **high**ly significant. Primitive **trib**es have an upbeat attitude.
aw	A large **hous**ehold needs lots of appliances. Every picture is worth a **thous**and words.
oy	The **boy** went to play tennis. Never exp**loit** the lives of the needy.

Table 1: The stimuli used in the experiment. The syllable in bold contains the diphthong join.

[3] An experimental version of rVoice was used in our research.

4.2 Test design

The sentences shown in table 1 were synthesised using an experimental version of the *rVoice* speech synthesis system using an adult North-American male voice. For each sentence we made a large number of synthetic versions, by varying the two diphone candidates which make the diphthong and keeping all the other units the same, as shown in figure 2. We have around 25 candidates for each diphone and thus obtained around 600 different synthetic versions.

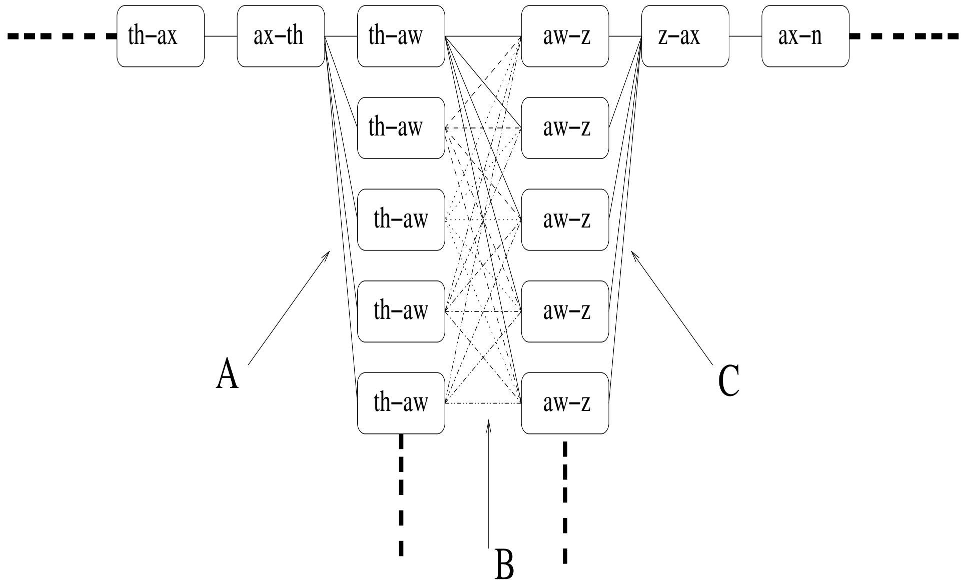

Figure 2: Synthesising the test stimuli

We removed those synthetic versions which had a prominent join in either the phone immediately before or the phone immediately after the the diphthong – marked as joins A and C in figure 2. The remaining versions were further pruned based on target features of the diphones making the diphthong, to ensure similar prosody among all synthetic versions. This process resulted in around 30 versions with a wide range of concatenation discontinuity at the diphthong join (B in figure 2), and all with good joins elsewhere (as perceived by the authors). We switched off smoothing to ensure no effect of it on concatenation discontinuities, though the quality of synthesiser is very high even without smoothing.

For each sentence in table 1, we selected what we judged to be the best and worst out of the 30 versions – these were used to set the end points of the listeners' rating scale, as described below.

4.3 Test procedure

There were around 25 participants in the listening test, most of them were graduate students with some experience of speech synthesis; most were native speakers of British English.

The test was carried out in blocks of around 35 test stimuli, with one block for each sentence in table 1. Listeners could take as long as they pleased over each block, and take a rest between blocks. We also gave a choice to listeners either to complete all the test blocks in same session or in few sessions, because of the difficulty involved of the task. Each test block contained a few duplications of some test stimuli to validate the listeners' scores.

At the start of each test block, listeners were first shown the written sentence with an indication of which word contains the join. Then they were presented with the two reference stimuli for that sentence: one containing the best and the other the worst join in order to set the end-points of a 1-to-5 scale. Listeners could listen to the reference stimuli as many times as they liked and they could also review them at regular intervals (for every 10 test stimuli) throughout the test.

Listeners were then played each test stimulus in turn and were asked to rate the quality of that join on a scale of 1 (worst) to 5 (best). They could listen to each test stimulus up to three times. Each test stimulus consisted of first the entire sentence, then only the word containing the join (extracted from the full sentence, not synthesised as an isolated word).

5 Results and Discussion

5.1 Listener Ratings

Consistency check

A validation set was included in the stimuli played to every listener; this consisted of a few duplications of some versions of the test stimuli. The size of this validation set is typically 6 out of 30 stimuli, thus making each test block of 36 stimuli. The listener scores can be cross checked using the validation set since identical stimuli are rated twice – we can compare the two ratings to measure listener consistency.

If the preference ratings for two identical stimuli are the same, then we give a score of 1, if the ratings are within 1 point on the 1-5 scale, we give a score of 0.5, otherwise we give a score of 0. For each listener, all scores for such validation stimuli were averaged and converted to a percentage. Listeners with consistency of less than 50% are discarded for each test block.

The above test will not catch listeners who consistently give the same or very similar ratings to *all* stimuli (e.g. everything is rated 3 out of 5). Such listeners were found by manual inspection of the data, and removed.

The mean listener scores reported below are only computed for the remaining listeners with more than 50% consistency. Table 2 shows the number of listeners for each sentence and the number of listeners with more than 50% consistency. The different number of listeners for each sentence is because some listeners did not complete the test before. The wide variation in number of consistent subjects per sentence may be due to the difficulty of the task; many listeners also commented that this was the case. For one sentence (the first sentence in *aw* row) all subjects were consistent, but we feel that this is mere coincidence. Also we did not observe any relation between consistency and correlations between listener scores and our distance measures.

5.2 Correlations with statistical distances

Correlation between mean listener scores for each synthetic version and join costs as calculated by a join cost function will be used to compare various join cost functions. As shown in the left-hand plot of figure 3, if the mean listener scores and join costs have a strong linear relationship, i.e. if we can fit a straight line with a positive slope through these points with

	no. of listeners	consistent listeners
ey	20, 20	14, 11
ow	19, 17	10, 9
ay	24, 17	12, 9
aw	17, 19	17, 11
oy	19, 20	10, 11

Table 2: Consistency of listeners in listening tests, each number in a pair corresponds to the two sentences for each diphthong listed in table 1.

a small mean squared error, then we can say that join cost and the perceptual scores are *highly correlated*. In contrast, as in the right-hand plot of figure 3, if a straight line cannot be fitted to these points with small error then they have *low correlation*. We compute the correlation coefficient, **r** using [22]:

$$\mathbf{r} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{11}$$

where, n is the number of pairs of measures, x and y. The value of **r** is always between -1.0 and +1.0. if the value of **r** is around 1.0, then there is a strong positive relation between two measures. Conversely, if the value correlation coefficient is around 0, then the two measures not correlated. A strong negative relationship is indicated by negative values of **r**. A good join cost function should have r values as close to 1 as possible.

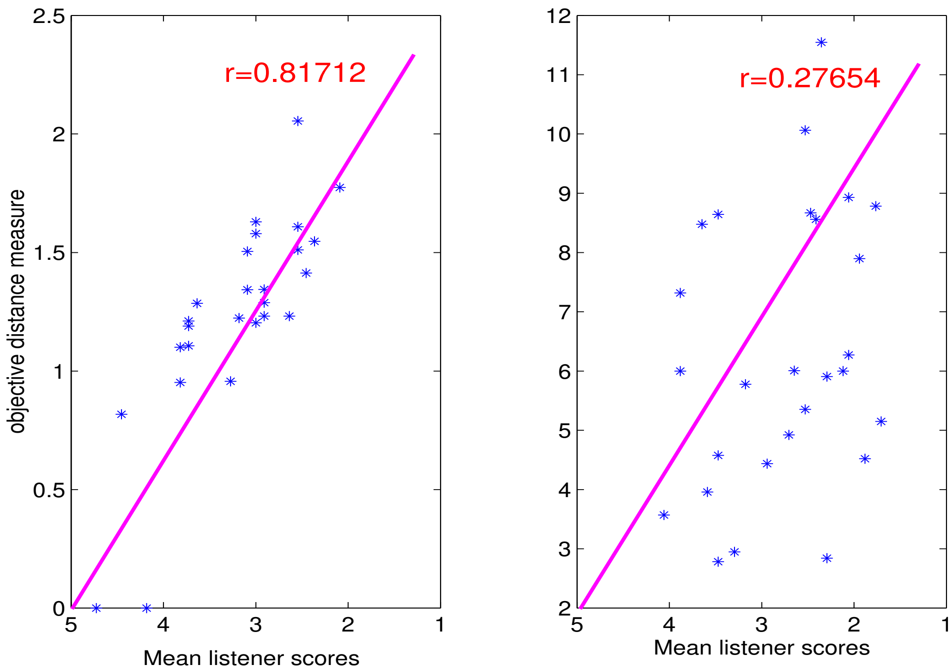

Figure 3: Example correlations

Once the correlations are computed, a significance test can be used to determine the probability that the observed correlation is a real one and not chance. A one-tailed test was chosen, since we know the direction of the relationship between the join cost and perceptual score: a low perceptual score – e.g. 1 out of 5 – rating corresponds to a high join cost. We computed the t statistic using [22]:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{12}$$

where n is the length of each test of block. Once we have the t statistic, we can refer to Student's t distribution table to find the significance (p) of the test.

	Euclidean		Absolute		Mahalanobis	
	MFCC	MFCC+Δ	MFCC	MFCC+Δ	MFCC	MFCC+Δ
ey	0.27	0.35	0.27	0.39	0.20	0.38
	0.65	**0.56**	**0.69**	**0.54**	**0.73**	**0.49**
ow	0.31	0.31	0.31	0.32	0.30	0.22
	0.57	**0.53**	**0.55**	**0.50**	**0.60**	**0.50**
ay	0.32	0.29	0.34	0.27	0.39	0.19
	0.61	**0.67**	**0.62**	**0.71**	**0.58**	**0.61**
aw	0.33	0.23	0.35	0.17	0.25	-0.05
	0.76	**0.76**	**0.73**	**0.74**	**0.78**	**0.75**
oy	0.08	0.06	0.08	0.06	0.24	0.24
	0.14	0.21	0.13	0.25	0.14	0.31

Table 3: Correlation between perceptual scores and various join costs using MFCCs.

Correlation coefficients of various spectral distance measures with mean listener preference ratings are reported in tables 3, 4 and 5. Correlation coefficients above the 1% significance level are in bold face.

It is immediately apparent that no distance measure performs well in all cases. The functions using LSFs have a higher number of significant correlations compared to those using MFCCs or MCAs. Unfortunately, none of these measures yield 1% significant level correlations for four of our ten sentences. Including delta coefficients, explained in section 3.1, did not generally improve correlations. Join cost functions using the symmetric KL distance have high correlations for more cases (five out of ten cases) than other distance measures, although a simple absolute distance measure performs as well as any other measure.

In table 3 it can be seen that Euclidean, absolute and Mahalanobis distance measures on MFCCs have good correlations with perceptual scores in many cases. Join cost functions using LSFs also have higher (i.e. stronger) correlations in more cases, as observed in table 4. From table 5, it is clear that join cost functions using MCA coefficients correlate well with perceptual scores in only a few cases compared to those using MFCCs and LSFs. This is contrary to our previous results [28], where we observed many 1% significant correlations. The results presented here are based on perceptual data from more listeners, and are therefore more reliable. Overall, we observed most 1% significant correlations for a Mahalanobis distance computed on LSFs plus delta coefficients.

However, MCAs do have an advantage over MFCCs and LSFs due to their compact size:

	Euclidean		Absolute		Mahalanobis		KL
	LSF	LSF+Δ	LSF	LSF+Δ	LSF	LSF+Δ	LSF
ey	0.04	0.06	0.13	0.20	0.30	0.39	0.30
	0.73	**0.73**	**0.75**	**0.75**	**0.75**	**0.68**	**0.73**
ow	0.40	0.36	0.35	0.25	0.31	0.14	0.33
	0.53	**0.54**	**0.46**	**0.47**	**0.48**	**0.47**	**0.45**
ay	0.16	0.14	0.12	0.07	0.23	0.07	0.38
	0.50	**0.58**	**0.51**	**0.64**	**0.56**	**0.62**	**0.63**
aw	0.21	0.27	0.09	0.26	0.19	**0.61**	0.16
	0.80	**0.80**	**0.79**	**0.80**	**0.81**	**0.80**	**0.79**
oy	0.18	0.20	0.11	0.15	0.13	0.27	0.15
	0.12	0.15	0.17	0.22	0.14	0.31	0.30

Table 4: Correlation between perceptual scores and various join costs using LSFs.

the size[4] of the MCA feature vector is only 12, whereas MFCCs are 26 and LSFs are 24 . Thus, using MCA coefficients reduces the memory required quite significantly. Considering this, the symmetric KL distance measure using MCA coefficients is also a good choice, which has four 1% significant correlations out of ten cases.

	Euclidean		Absolute		Mahalanobis		KL
	MCA	MCA+Δ	MCA	MCA+Δ	MCA	MCA+Δ	MCA
ey	0.36	0.29	0.34	0.34	0.36	0.40	**0.45**
	0.62	**0.49**	**0.62**	**0.49**	**0.58**	**0.63**	**0.65**
ow	0.02	0.12	0.06	0.14	0.13	0.06	0.11
	0.39	**0.48**	0.40	**0.50**	**0.47**	**0.44**	0.32
ay	-0.02	0.06	-0.03	0.00	0.01	0.03	0.11
	0.41	0.22	0.38	0.23	0.39	0.39	**0.45**
aw	**0.46**	0.28	0.36	0.33	0.35	0.29	0.36
	0.76	**0.54**	**0.75**	**0.54**	**0.79**	**0.71**	**0.82**
oy	0.28	**0.51**	0.24	**0.51**	0.17	0.21	0.16
	0.12	0.24	0.14	0.37	0.20	0.28	0.27

Table 5: Correlation between perceptual scores and various join costs using MCA coefficients.

5.3 Correlations with weighted distances

In order to improve correlations further, we constructed various join cost functions using a weighted sum of sub-costs. We tried two types of weighted functions: one is a weighted sum of join costs from above, the other applies weighting to individual MCA coefficients. The weights were computed using a least-squares approach, as explained in section 3.4. These results are different from our previously published results [27] due to an increased amount perceptual data.

[4]all these figures include delta coefficients

Weighted sums of join costs

Table 6 summarises correlations for absolute distances using MFCCs, LSFs or MCA coefficients and for a join cost using a weighted sum of these three sub-costs. The weights used for MFCCs, LSFs and MCA coefficients were 0.11, 0.05, 0.84 respectively. These weights were found using the whole data set (i.e. closed test).

	MFCC	LSF	MCA	weighted dist.
ey	0.27	0.14	0.34	0.35
	0.69	**0.75**	**0.62**	**0.70**
ow	0.31	0.35	0.06	0.21
	0.55	**0.46**	0.40	**0.51**
ay	0.34	0.12	-0.03	0.20
	0.62	**0.51**	0.38	**0.58**
aw	0.35	0.09	0.36	**0.44**
	0.73	**0.79**	**0.75**	**0.76**
oy	0.08	0.11	0.24	0.17
	0.13	0.17	0.14	0.13

Table 6: Correlation between perceptual scores and absolute distances using MFCCs, LSFs or MCAs and weighted sum of the three distances.

The correlation coefficients of Euclidean distance measures of all three spectral features and their weighted sum with mean listener ratings are reported in table 7. The weights used were 0.07 (MFCCs), 0.00(LSFs) and 0.93(MCAs).

	MFCC	LSF	MCA	weighted dist.
ey	0.27	0.04	0.36	0.34
	0.65	**0.73**	**0.62**	**0.68**
ow	0.31	0.40	0.01	0.20
	0.57	**0.53**	0.39	**0.52**
ay	0.33	0.16	-0.02	0.22
	0.62	**0.50**	0.41	**0.58**
aw	0.43	0.21	**0.46**	**0.46**
	0.76	**0.80**	**0.76**	**0.78**
oy	0.07	0.18	0.28	0.17
	0.14	0.12	0.11	0.13

Table 7: Correlation between perceptual scores and Euclidean distances using MFCCs, LSFs, MCA coefficients and weighted sum of the three distances.

From table 7 it is evident that we can improve correlations by setting weights on individual distances, e.g MFCCs produce good correlations for *ay*, MCAs yield better correlations for *aw*, and a weighted sum achieves good correlations for both cases. Thus, we achieve a 1% significant correlation if more than one of the distances of three component costs have a 1% significant correlation. In table 8 we present correlations between perceptual scores and Mahalanobis distances of MFCCs, LSFs and MCA coefficients and their weighted sum.

	MFCC	LSF	MCA	weighted dist.
ey	0.20	0.30	0.36	0.33
	0.73	**0.75**	**0.58**	**0.69**
ow	0.30	0.31	0.13	0.22
	0.60	**0.48**	**0.47**	**0.56**
ay	0.39	0.23	0.01	0.19
	0.58	**0.56**	0.39	**0.51**
aw	0.25	0.19	0.35	0.36
	0.78	**0.81**	**0.79**	**0.81**
oy	0.24	0.13	0.17	0.24
	0.14	0.14	0.20	0.18

Table 8: Correlation between perceptual scores and Mahalanobis distances of MFCCs, LSFs, MCA coefficients and weighted sum of above three measures.

The weights used are 0.39(MFCCs), 0.0(LSFs) and 0.61(MCAs). Results are not as good as for the Euclidean distance measure.

Weighting individual MCA coefficients

In the weighted sum experiments, we observe that MCA coefficients have higher weights than MFCCs or LSFs. Also, as we previously mentioned, the size of the MCA feature vector is small and therefore attractive from a memory usage point of view. Hence, we carried out a further experiment in which the individual MCA coefficients were weighted. The least-squares method did not yield good solutions in this case, so instead we randomly generated sets of weights and chose the sets which produced the highest number of 1% significant correlations. This again was a closed-set experiment, but in future we compute the weights using an open set experiment. parameters of MCA.

MCA parameter	set1	set2	set3
F1	0.128	0.048	0.104
F2	0.145	0.031	0.111
F3	0.025	0.007	0.035
B1	0.109	0.227	0.138
B2	0.128	0.180	0.088
B3	0.077	0.046	0.125
E1	0.038	0.062	0.062
E2	0.008	0.017	0.008
E3	0.014	0.009	0.013
DF1	0.018	0.099	0.118
DF2	0.114	0.111	0.071
DF3	0.196	0.163	0.127

Table 9: Three sets of weights used on individual MCA coefficients: formant frequency (F), bandwidth (B), energy (E), delta-formant frequency (DF).

Table 9 shows three different sets of weights on MCA coefficients, and the corresponding correlations obtained are shown in table 10. Sets 2 and 3 produced seven 1% significant correlations out of ten cases, and even achieved good correlations for the *oy* diphthong, which previously had very poor correlations with all other distance measures (see tables 6, 7 and 8).

	set1	set2	set3
ey	**0.46**	**0.48**	**0.45**
	0.69	**0.67**	**0.70**
ow	0.12	0.07	0.17
	0.58	**0.56**	**0.58**
ay	0.01	0.06	0.02
	0.25	0.18	0.23
aw	0.39	**0.47**	**0.48**
	0.63	**0.64**	**0.62**
oy	**0.51**	**0.47**	**0.49**
	0.54	**0.56**	**0.46**

Table 10: Correlation between perceptual scores and absolute distances based on weighted MCA coefficients.

6 Conclusions

Finding a join cost function whose output correlates well with human listeners' perception of join discontinuity is difficult. The correlations between perceptual data and various join costs based on three speech parameterisations – Mel frequency cepstral coefficients (MFCCs), line spectral frequencies (LSFs) and multiple centroid analysis (MCA) coefficients – and four distance measures – Euclidean, absolute, Mahalanobis and symmetric Kullback-Leibler (KL) – were computed. The correlation results suggest that Mahalanobis distance using LSFs plus delta coefficients is a reasonable join cost. If we consider the storage requirements of the inventory then KL distance using MCAs is also a reasonable choice.

6.1 Weighted sums of join costs

In order to achieve a higher number of 1% significant correlations in our ten sentences (as listed in table 1), we constructed new join costs as weighted sums of individual join costs. However, this weighting did not result in much improvement, and of course has much higher storage requirements since multiple parameterisations of the speech signal must be stored. To solve this problem, a second type of weighted join cost was constructed in which individual MCA coefficients were weighted. This join cost gave the most promising results: seven 1% significant correlations out of ten cases.

6.2 The listening test

We used a perceptual listening test to measure the degree of perceived concatenation discontinuity in synthetic speech. The stimuli were created by a typical unit-selection synthesiser and therefore we believe the results we obtain on such stimuli apply to a) any speech produced by the same synthesiser, not just American English diphthongs, and b) to any other typical unit-selection synthesiser. Also, the stimuli we used are natural sentences, which take into account co-articulation and wider contexts.

6.3 Correlation as an evaluation tool

Since we have a 1-5 scale for rating the discontinuities, correlation is used as an evaluation tool. First, we computed correlations of individual perceptual data (i.e. each of our ten cases) with various distance measures. Then obtained statistical significance of these correlation values. We also computed a global correlation by pooling all the perceptual data. This is low (around 0.3 for weighted distances) but it is statistically very significant. These values are high for weighted MCAs (0.31,0.33, 0.35 for corresponding set of weights in table 9) and thus re-confirm their high prediction rates.

6.4 Future work

The most obvious extension to our work would be to use a wider variety of speech segments than just diphthongs, and to use a variety of "voices", in multiple languages.

The computation of join cost and spectral smoothing are intimately related. Suppose that we had a sufficiently large database and a perfect measure of join cost: no smoothing would be required. Conversely, if we could smooth joins better, then the method of computing join cost would be less critical, and perhaps a smaller inventory would be needed. Intuitively, if we combine the two operations of join cost calculation and join smoothing we should be able to obtain better results than by considering them independent processes. In our ongoing work, we are investigating the use of a single representation and/or model of the speech signal, which can be used for join cost computation as well as smoothing. In this way, we hope that when the join cost is calculated as low that means either a join is already good, or can be made good by smoothing, and if a join cost is calculated as high, that means that no amount of smoothing will be able to make it sound good.

Acknowledgements

Thanks to Rhetorical Systems Ltd.[5] for the use of rVoice and for partially funding the first author. Thanks also to all the experimental subjects: the members of CSTR, staff at Rhetorical Systems Ltd. and students on the M.Sc. in Speech and Language Processing, University of Edinburgh. The authors also acknowledge the assistance of Dr. Alice Turk of the dept. of Theoretical and Applied Linguistics in designing the listening tests.

References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen TTS system. In *Proc. Joint Meeting of ASA, EAA, and DEGA*, Berlin, Germany, 1999.

[5] www.rhetorical.com

[2] C. Blouin, O. Rosec, P. Bagshaw, and C. d'Alessandro. Concatenation cost calculation and optimisation for unit selection in TTS. In *Proc. IEEE 2002 WorkShop on Speech Synthesis*, Santa Monica, USA, September 2002.

[3] N. Campbell. CHATR: A high-definition speech re-sequencing system. In *Proc. 3rd ASA/ASJ Joint Meeting*, pages 1223–1228, Hawaii, 1996.

[4] Jing-Dong. Chen and N. Campbell. Objective distance measures for assessing concatenative speech synthesis. In *Proc. Eurospeech*, Budapest, Hungary, 1999.

[5] Min Chu and Hu Peng. An objective measure for estimating MOS of synthesized speech. In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[6] A. Crowe and M.A. Jack. Globally optimising formant tracker using generalised centroids. *Electronic Letters*, 23(19):1019–1020, 1987.

[7] R.E. Donovan and E.M Eide. The IBM trainable speech synthesis system. In *Proc. ICSLP*, Sydney, Australia, 1998.

[8] Robert E. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesisers. In *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 59–62, Pethshire, Scotland, 2001.

[9] M. Founda, G. Tambouratzis, A. Chalamandaris, and G. Carayannis. Reducing spectral mismathces in concatenative speech synthesis via sysematic database enrichment. In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[10] X. Huang, A. Acero, and Hon Hasio-wuen. *Spoken Language Processing, A guide to Theory , Algorithms and Sysystem Development.* Prentice Hall, 2001.

[11] A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, pages 373–376, 1996.

[12] H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative synthesis. In *ICSLP*, Denver, USA, 2002.

[13] E. Klabbers and R. Veldhuis. On the reduction of concatenation artefacts in diphone synthesis. In *Proc. ICSLP*, volume 6, pages 1983–1986, Sydney, Australia, 1998.

[14] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Trans. Speech and Audio Processing*, 9(1):39–51, 2001.

[15] E. Klabbers, R. Veldhuis, and K. Koppen. A solution to the reduction of concatenation artefacts in speech synthesis. In *Proc. ICSLP*, Beijing, China, 2000.

[16] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[17] Minkyu Lee. Perceptual cost functions for unit searching in large corpus-based concatenative text-to-speech. In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[18] J. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English Speech: A Dynamic Approach.* Springer, 1993.

[19] R.D. Patterson, M.H. Allerhand, and C. Giguere. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform". *J. Acoust. Soc. Am.*, 98:1890–1894, 1995.

[20] H. Peng, Y. Zhao, and M. Chu. Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation. In *Proc. ICSLP*, Denver, USA, 2002.

[21] F.K. Soong and B.H. Juang. Line spectrum pairs (LSP) and speech data compression. In *Proc. ICASSP*, pages 1.10.1–1.10.4, 1984.

[22] Murray R. Spiegel. *Theory and Problems of Probability and Statistics*. Schaum's Outline Series in Mathematics. McGraw-Hill Book Company, 1975.

[23] Y. Stylianou and Ann K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proc. ICASSP*, Salt Lake City, USA, 2001.

[24] Ann K. Syrdal. Phonetic effects on listener detection of vowel concatenation. In *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[25] M. Tsuzaki and H. Kawai. Feature extraction of unit selection in concatenative speech synthesis: Comparison between AIM, LPC, and MFCC. In *ICSLP*, Denver, USA, 2002.

[26] R. Veldhuis and E. Klabbers. On the computation of Kullback-Leibler measire for spectral distances. *IEEE Trans. Speech and Audio Processing*, 11(1):100–103, 2003.

[27] J. Vepa, S. King, and P. Taylor. New objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proc. IEEE 2002 WorkShop on Speech Synthesis*, Santa Monica, USA, September 2002.

[28] J. Vepa, S. King, and P. Taylor. Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *ICSLP*, Denver, USA, 2002.

[29] J. Wouters and M. Macon. Perceptual evaluation of distance measures for concatenative speech synthesis. In *Proc. ICSLP*, volume 6, pages 2747–2750, Sydney, Australia, 1998.

[30] A.A. Wrench. Analysis of fricatives using multiple centres of gravity. In *Proc. International Congress of Phonetic Sciences*, volume 4, pages 460–463, 1995.