IDIAP RESEARCH REPORT

Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis

Jithendra Vepa [a] Simon King [b]

IDIAP–RR 05-34

June 2005

to appear in
IEEE Transactions on Speech and Audio Processing

[a] IDIAP Research Institute, P. O. Box 592, Rue du Simplon 4, 1920 Martigny, Switzerland.
[b] Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK.

IDIAP Research Report 05-34

Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis

Jithendra Vepa Simon King

June 2005

to appear in
IEEE Transactions on Speech and Audio Processing

Abstract. In unit selection-based concatenative speech synthesis, *join cost* (also known as concatenation cost), which measures how well two units can be joined together, is one of the main criteria for selecting appropriate units from the inventory. Usually, some form of local parameter smoothing is also needed to disguise the remaining discontinuities. This paper presents a subjective evaluation of three join cost functions and three smoothing methods. We also describe the design and performance of a listening test. The three join cost functions were taken from our previous study, where we proposed join cost functions derived from spectral distances, which have good correlations with perceptual scores obtained for a range of concatenation discontinuities. This evaluation allows us to further validate their ability to predict concatenation discontinuities. The units for synthesis stimuli are obtained from a state-of-the-art unit selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd. In this paper, we report listeners' preferences for each join cost in combination with each smoothing method.

1 Introduction

Unit selection-based concatenative speech synthesis systems [1, 2, 3, 4] have become popular recently because of their highly natural-sounding synthetic speech. These systems have large speech databases containing many instances of each speech unit (e.g. diphone), with a varied and natural distribution of prosodic and spectral characteristics. When synthesising an utterance, the selection of the best unit sequence from the database is based on a combination of two costs: target cost (how closely candidate units in the inventory match the required targets) and join cost (how well neighbouring units can be joined) [1]. The target cost is calculated as the weighted sum of the differences between the various prosodic and phonetic features of target and candidate units. The join cost, also known as concatenation cost, is also determined as the weighted sum of sub-costs, such as absolute differences in F0, amplitude and mismatch in spectral (acoustic) features. The optimal unit sequence is then found by a Viterbi search for the lowest cost path through the lattice of the target and concatenation costs.

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 perception of discontinuity at unit concatenation points. In other words: the perfect join cost should predict the degree of perceived discontinuity.

A few recent studies have attempted to determine which objective distance measures are best able to predict audible concatenation discontinuities. Klabbers and Veldhuis [5] examined various distance measures on five Dutch vowels to reduce the concatenation discontinuities in diphone synthesis and found that the Kullback-Leibler measure on LPC power normalised spectra was the best predictor. A similar study by Wouters and Macon [6] for unit selection, showed that the Euclidean distance on Mel-scale LPC-based cepstral parameters was a good predictor, and utilising weighted distances or delta coefficients could improve the prediction. Stylianou and Syrdal [7] found that the Kullback-Leibler distance between FFT-based power spectra had the highest detection rate. Donovan [8] proposed a new distance measure which uses a decision-tree based context dependent Mahalanobis distance between perceptual cepstral vectors.

All these previous studies focused on human detection of audible discontinuities in **isolated words** generated by concatenative synthesisers. We extended this work to the case of **polysyllabic words in natural sentences** and new spectral features, multiple centroid analysis (MCA) coefficients [9, 10]. We designed and conducted a perceptual experiment to measure the correlations between mean listener ratings and various join costs and reported the results in [11, 12, 13, 14].

In this study, we have designed another listening test to evaluate the best three join cost functions obtained from our previous perceptual experiments. This test is to further validate their ability to predict concatenation discontinuities. Each of the three join cost functions is combined with each of three different smoothing methods, including a novel Kalman filter-based method. The listening test is also intended to discover whether the smoothed line spectral frequencies (LSFs) obtained from the Kalman filter produce better synthesis than LSFs smoothed by other methods. We used our own implementation of residual excited linear prediction (RELP) synthesis for waveform generation using units selected by the *rVoice* synthesis system from Rhetorical Systems Ltd.[1]

In the next section we briefly describe our previous perceptual listening experiment and various spectral distance measures used in join cost functions. In section 3, we discuss different smoothing techniques evaluated in this paper. Also, we explain the implementation of the RELP synthesis method used for waveform generation. In section 4, we discuss the design and procedure of the listening test. Finally, we present subjective results of these various combinations and discuss them in section 5.

[1]We did not use *rVoice* for waveform generation as we have no access to its source code and can only plug-in join cost code.

2 Join cost functions

We have chosen three of the best spectral distances from our previous studies [11, 12, 13, 14] based on the number of statistically significant correlations with data obtained from perceptual experiment. For clarity of the reader, here we briefly explain the design of our perceptual experiment and various spectral distance measures used in join cost functions.

2.1 Perceptual Listening Experiment

A listening test was designed to measure the degree of **perceived** concatenation discontinuity in natural sentences generated by the state-of-the-art speech synthesis system, *rVoice* using an adult North-American male voice. We focused on diphthong joins where spectral discontinuities are particularly prominent due to moving formant values. We selected two natural sentences for each of five American English diphthongs (ey, ow, ay, aw and oy) [15]. One word in each 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.

These sentences were then synthesised using the experimental version of *rVoice* speech synthesis system. For each sentence we made various synthetic versions, by varying the two diphone candidates which make the diphthong and keeping all the other units the same. We pruned several synthetic versions based on joins of neighbouring units and prosodic features of diphones making the diphthong. This process resulted in around 30 versions with variation in concatenation discontinuities at the diphthong join. The authors manually selected what they judged to be the best and the worst synthetic versions by listening to these 30 versions. This process was repeated for each sentence in Table 1.

There were 35 participants in our perceptual listening test, most of them were native speakers of British English with some experience of speech synthesis. Subjects were first shown the written sentence, with an indication of which word contains the join. At the start of the test they were first presented with a pair of reference stimuli: one containing the best and the other the worst joins (as selected by the authors) in order to set the endpoints of a 1-to-5 scale. They can listen to reference stimuli as many times as they liked. They 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).

2.2 Spectral Distance Measures

We used three parameterisations: Mel Frequency Cepstral Coefficients (MFCCs) [16], Line Spectral Frequencies (LSFs) [17, 18] and Multiple Centroid Analysis (MCA) coefficients [9, 10]. Standard distance measures: Euclidean, absolute, Kullback-Leibler and Mahalanobis distances were computed for all the above speech parameterisations.

We investigated many different ways to compute spectral distance measures to use in join cost functions. First, we computed a simple single-frame distance, i.e. using only the final frame of the first unit and the initial frame of the second unit. Then, we extended to multi-frame distances, where we used several frames of the two units to compute the distance. Our preliminary observations of correlations of join cost functions with single-frame distances indicated that proper weighting of various distance metrics and speech parameters can improve the correlations further. This lead to our investigations on combining distance metrics, speech parameterisations and multi-frame distances [12].

A probabilistic approach for join cost computation was proposed in [13], which uses a linear dynamic model (LDM)[2], sometimes known as **Kalman filters** [20], to model line spectral frequency trajectories. The model uses an underlying subspace in which it makes smooth, continuous trajectories. This subspace can be seen as an analogy for underlying articulatory movement. Once trained, the model can be used to measure how well concatenated speech segments join together. The objective join cost is based on the error between model predictions and actual observations, computed from the log likelihood of the observation sequence given the model. We experimented with three models which differed in initial conditions and three analytical measures which are derived from the shape of negative log likelihood curve.

2.3 Correlation Results

We computed correlations between mean listener scores obtained from perceptual experiments and various spectral distance measures used in join cost functions. Then, we observed out of our 10 cases (i.e., 10 sentences in table 1), how many times the spectral distance measures produced 1% significant correlations[3]. The main focus of significance of correlations is to generalise the distance measure for all the phones. Though we tested distance measures only on diphthong joins, our hypothesis is that if a distance measure or join cost function works for diphthongs, which have difficult joins, then it will perform well for other phones. Furthermore, the join cost functions that perform well on a large number of cases of diphthongs are expected to generalise better to other phone classes.

As we mentioned earlier in this section, we have chosen three of the best spectral distance measures from among those described in 2.2 based on number of 1% significant correlations. Three spectral distance measures and our names for the join cost functions derived from them are as follows:

1. *Mahalanobis distance on line spectral frequencies (LSF) and their deltas of frames at the join. The join cost function based on this is termed **LSF join cost**.*

2. *Mahalanobis distance computed using multiple centroid analysis (MCA) coefficients of multi-frames (seven frames, i.e. three frames on either side of join plus one frame at the join). The join cost function based on this is termed **MCA join cost**.*

3. *The join cost derived from the negative log likelihood estimated by running the Kalman filter on LSFs of the phone at the join is termed **Kalman join cost**.*

The first join cost function above scored **six** 1% significant correlations out of a possible maximum of 10. There were **seven** 1% significant correlations for the second measure and **five** for the third. The rankings of these three join costs are therefore as shown in table 2.

[2]LDMs can also be used for speech recognition [19].
[3]Correlations at p-value < 0.01 significance

Rank	*Join Cost*
1	MCA join cost
2	LSF join cost
3	Kalman join cost

Table 2: Rankings for three join costs, obtained in the first listening test

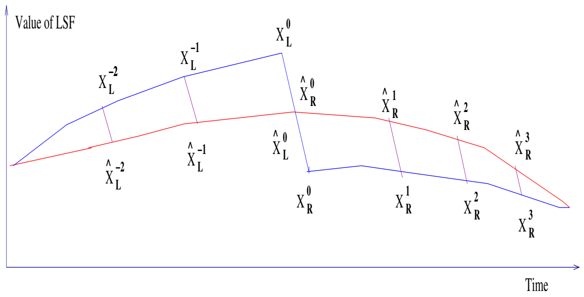

Figure 1: Linear smoothing on parameters (LSFs) of frames at the join (adapted from [22]).

3 Smoothing techniques

After units are concatenated, most systems attempt some form of local parameter smoothing to disguise the remaining discontinuity. One of our goals is to combine the join cost function and the join smoothing process in some optimal way as these two operations interact closely. Suppose, a large database and a perfect join cost function are available then no smoothing would be required. On the other hand, the join cost function would be less important if we could smooth joins better.

a) No Smoothing: In this case, we do not perform any smoothing on the spectral features or on the resulted speech signal from RELP synthesis.

b) Linear smoothing: The line spectral frequencies have good interpolation properties and yield stable filters after interpolation [21]. Although LSF interpolation is widely used in speech coding, it can also be used for speech synthesis. Dutoit [22] showed that LSFs have good interpolation properties and produce smoother transitions than LPC parameters. LSF interpolation was compared with other smoothing methods in [23] and performed well in many cases.

We have implemented linear smoothing on LSFs of a few frames of the phones at the join as presented in [22]. The main idea of this technique is to distribute the difference of the LSF vectors at the join across a few frames on either side of the join. To explain this technique, consider L and R as left and right segments at the join and X is a LSF vector. Assume the number of frames on the left side and the right side of the join to be M_L and M_R respectively. Then, the LSFs after smoothing ($\hat{X}$) are:

$$\hat{X}_L^{-i} = X_L^i + (X_R^0 - X_L^0)\,\frac{M_L - i}{2M_L} \qquad 0 \le i < M_L \tag{1}$$

$$\hat{X}_R^{j} = X_R^j + (X_L^0 - X_R^0)\,\frac{M_R - j}{2M_R} \qquad 0 \le j < M_R \tag{2}$$

where X_L^0 and X_R^0 are frames at the end of L and beginning of R, i.e. exactly at the join. The function of linear smoothing is showed in figure 1, where M_L and M_R are 2 and 3 respectively.

c) Kalman filter-based smoothing: Linear dynamic models, which are used to compute the Kalman join cost functions [13], can *also* smooth the observations (LSFs in our case) since running a Kalman filter involves computing the most likely (smoothed) observations. These smoothed LSFs are then used in RELP synthesis to generate the synthetic waveform. We investigate the combined Kalman filter-based join cost function and Kalman smoothing operation as one possible approach towards the above objective. So, in the listening test, we also directly compare the Kalman smoothing operation to the linear smoothing technique.

3.1 Residual excited linear prediction (RELP) based synthesis

Residual excited LP (RELP) is one of the standard methods for resynthesis, which is also used in Festival [24]. In this method, first LPC analysis has to be carried out on the original speech to obtain LPC parameters. Then, inverse filtering is performed to get the residual signal. Consider original speech sample $x[n]$ which can be predicted as a linear combination of the previous p (linear prediction order) samples, as given below:

$$\tilde{x}[n] = \sum_{i=1}^{p} -a_i \; x[n-i] \tag{3}$$

where a_i are prediction coefficients and $x[n-i]$ are past speech samples. The prediction error due to this approximation is:

$$e[n] = x[n] - \tilde{x}[n] = x[n] + \sum_{i=1}^{p} a_i \; x[n-i] \tag{4}$$

This error is known as the *residual signal*, which can be used as the excitation to the LPC filter to get a perfect reconstruction of the speech signal.

During LPC analysis we have computed the LPC parameters using asymmetric[4] Hanning-windowed pitch-synchronous frames of the original speech as shown in figure 2. The advantage of using the asymmetric window can be observed in the figure, where successive pitch periods are very different in size and the window is not centered. The sample plots shown in the figure are two pitch periods in length. The residual is computed by passing the windowed original speech (plot (c)) through the inverse LPC filter. A sample residual signal is depicted in plot (d) of the figure 2.

Once the units are selected using the *rVoice* synthesis system, the corresponding LPCs and residual signals from the database are assembled. We convert the LPC parameters to LSFs, then employ one of two smoothing methods (linear or Kalman filter-based) and then convert back to LPC parameters for synthesis. The residual is not modified by the smoothing operation. Then, the LPC filter is excited using the residual to reconstruct the output speech waveform. In figure 2, the output waveform is depicted in the last plot, which is a reconstruction of the original signal. To get the full synthetic waveform for an utterance we overlap and add these two-pitch-period output waveforms.

4 Listening test

A listening test was designed to evaluate the three join costs and the above smoothing methods, and to compare the smoothed LSFs obtained from Kalman filter and linear smoothing on LSFs. We are testing the following three things:

- *Compare three join costs: LSF join cost, MCA join cost and Kalman join cost, irrespective of smoothing methods*
- *Similarly, compare three smoothing methods: no smoothing, linear smoothing and Kalman smoothing, irrespective of join cost.*
- *Check if Kalman join cost together with Kalman smoothing is any better than LSF join cost with linear smoothing.*

[4]The left and right halves of the window are different.

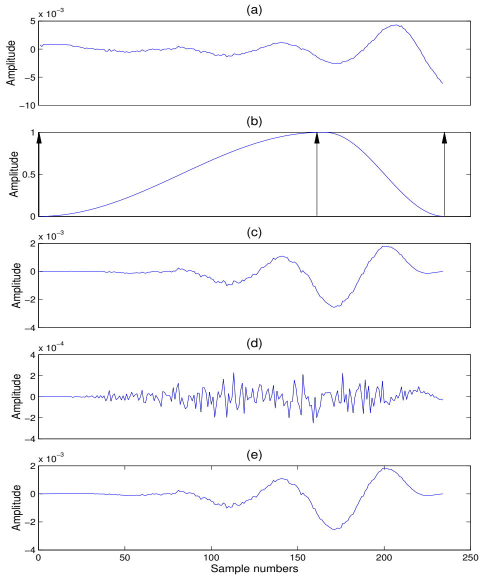

Figure 2: RELP synthesis using an asymmetric window: (a) Original waveform (b) Asymmetric Hanning window (pitch marks shown as arrows) (c) Windowed original waveform (d) Residual signal (e) Reconstructed waveform

4.1 Test design & stimuli

To describe our test design, we use 1, 2 and 3 to denote the three join costs: LSF, MCA and Kalman respectively. The three smoothing methods: a, b and c are no smoothing, linear smoothing and Kalman smoothing in that order. Now, we have 9 different synthetic versions for each of our test sentences obtained with the three join costs and the three smoothing methods, for example V_{1a} means synthesised version using join cost function "1" and smoothing method "a".

Ideally, to know which combination of join cost and smoothing method is the best, we need to compare all the combinations from 9 different versions. Such combinations formed from 9 versions result in 36 pairs[5], as shown in table 3, which are divided into 12 symmetric[6] blocks.

To know which join cost performs better, the three blocks in the first row need to be considered. Similarly, to compare smoothing methods three blocks in the second row have to be taken. The remaining two rows (in addition to first and second rows) are required to know which particular join cost and smoothing pair performs better than any other possible pair. However, this increases the number of our test stimuli and it is then not possible to test on many sentences.

In other words, if we consider all 36 pairs, a maximum of four sentences can be tested assuming the test duration is 30-40 minutes. In addition, subjects may lose interest after listening to the same sentences many times. To avoid the latter problem, we can rotate the various blocks between different subjects, i.e. presenting only a few (say 3 out of 12) blocks of each sentence and thus increasing the number of sentences to each subject. But in this case, we will not get as many subjective results per sentence because 4 subjects are used to test one sentence.

Hence we compared only one pair in the last two rows: Kalman join cost and Kalman smoothing *vs* LSF join cost and linear smoothing (i.e. V_{3c} vs V_{1b}). We have chosen linear smoothing since it is a popular and standard procedure in current synthesis systems and we feel combining this with one of our best join costs, the *LSF join cost*, becomes a strong contestant to V_{3c}. To do this comparison we added the V_{3c} and V_{1b} pair in our test stimuli to the first two rows of table 3.

[5]Each pair means one comparison, for example $V_{1a} - V_{2a}$

[6]Each block has an equal number of a particular version, for example in the first block V_{1a} appears twice, similarly V_{2a} and V_{3a} appear twice.

V_{1a}-V_{2a} V_{2a}-V_{3a} V_{3a}-V_{1a}	V_{1b}-V_{2b} V_{2b}-V_{3b} V_{3b}-V_{1b}	V_{1c}-V_{2c} V_{2c}-V_{3c} V_{3c}-V_{1c}
V_{1a}-V_{1b} V_{1b}-V_{1c} V_{1c}-V_{1a}	V_{2a}-V_{2b} V_{2b}-V_{2c} V_{2c}-V_{2a}	V_{3a}-V_{3b} V_{3b}-V_{3c} V_{3c}-V_{3a}
V_{1a}-V_{2b} V_{2b}-V_{3c} V_{3c}-V_{1a}	V_{2a}-V_{3b} V_{3b}-V_{1c} V_{1c}-V_{2a}	V_{3a}-V_{1b} V_{1b}-V_{2c} V_{2c}-V_{3a}
V_{1a}-V_{2c} V_{2c}-V_{3b} V_{3b}-V_{1a}	V_{2a}-V_{3c} V_{3c}-V_{1b} V_{1b}-V_{2a}	V_{3a}-V_{1c} V_{1c}-V_{2b} V_{2b}-V_{3a}

Table 3: All possible pairwise comparisons

The test sentences used in our listening test are presented in table 4. These eight sentences were selected randomly from twenty such sentences.

Sentence 1	Paragraphs can contain many different kinds of information.
Sentence 2	The aim of argument, or of discussion, should not be victory, but progress.
Sentence 3	He asked which path leads back to the lodge.
Sentence 4	The negotiators worked steadily but slowly to gain approval for the contract.
Sentence 5	Linguists study the science of language.
Sentence 6	The market is an economic indicator.
Sentence 7	The lost document was part of the legacy.
Sentence 8	Tornadoes often destroy acres of farm land.

Table 4: Listening test sentences

4.2 Test procedure

The listening test is divided into two parts to provide a few minutes break for the subjects. Each part consists of 96 pairs of synthetic stimuli covering the pairs in all blocks of the first two rows in the table 3, including one pair ($V_{3c} - V_{1b}$) and some validation pairs, i.e. presenting the above pairs in reverse order ($V_{1b} - V_{3c}$).

In each part, the two rows including a pair ($V_{3c} - V_{1b}$) and two validation pairs are presented alternatively to each subject as shown in figure 3. In this figure R1 and R2 each consist of 12 pairs of synthetic stimuli and covered in two parts (PART1 and PART2) for 8 sentences. The pairs for all sentences were randomised within each part of the test and presented to the subjects. For each pair of stimuli they are asked to judge which one is better by keying 1 or 2. This is a forced choice.

There were 33 participants in this listening test. Most of them were people in CSTR or students in the Dept. of Linguistics with some experience of speech synthesis. Around half of them were native speakers of British English. The tests were conducted in sound-proof booths using headphones. After the first part, the subjects were asked to take a rest for a few minutes. On the average, each part took

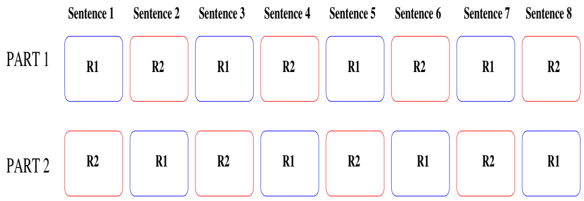

Figure 3: Test procedure, in each part the two rows (R1 and R2) are presented alternatively.

around 15 minutes and about 30-40 minutes for completion of two parts. The informal feedback from the subjects indicated that there was not much difference between the two stimuli in many pairs. In fact, a few of them felt that those pairs were the same, hence found it a difficult task.

4.3 Validation procedures

We have designed a couple of validation procedures to validate subjects' scores and to check the consistency of the subjects. These procedures will catch those subjects who give random scores in any part of the test, which are described below:

To check the validity of the subjects' results, we included 16 validation pairs in each part of the test. These pairs appear in reverse order. We adopted a scoring system where subjects are given a score of 1 or 0 for each of these 16 pairs. If subjects keyed the same response (i.e. 1 or 2) for the original pair and the validation pair then it is an error and they get a score of 0 as they preferred different stimuli in the original to the validation pair. If they key opposite responses (for example, 1 for original pair and 2 for validation pair) then they will get a score of 1. These scores are accumulated for 16 pairs for each part of the test. In figure 4, we have shown the number of parts which have equal or more validation scores for each validation cutoff ranging from 1 to 16. For example, the number 37, on top of the bar corresponding to the validation cutoff 10, indicates the number of parts which got a validation score of 10 or more.

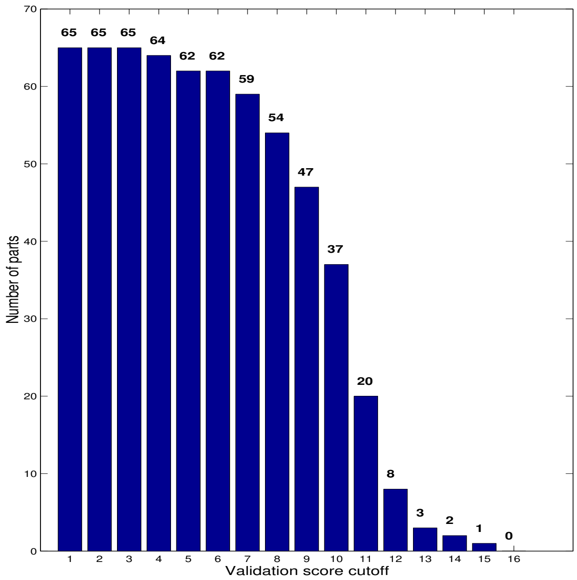

Figure 4: Subjects validity, number of parts with equal or more validation scores for validation cutoffs from 1 to 16.

We performed another validation procedure on the block level. Consider the first block in table

3; $V_{1a} - V_{2a}$, $V_{2a} - V_{3a}$ and $V_{3a} - V_{1a}$. If subjects preferred all the first stimuli (V_{1a}, V_{2a} and V_{3a}) then the block becomes invalid because, if they prefer V_{1a} and V_{2a}, then for the third pair, the valid selection is V_{1a}. Similarly, they can not prefer all the second stimuli in a block.

5 Subjective evaluation

5.1 Join costs

In figure 5, we show preferences for the three join costs for each sentence using the subjects who got validation scores of 10 or more out of 16 after removing invalid blocks. It can be observed from the figure that LSF join cost is preferred more times than MCA join cost and Kalman join cost. The Kalman join cost has the least number of preferences.

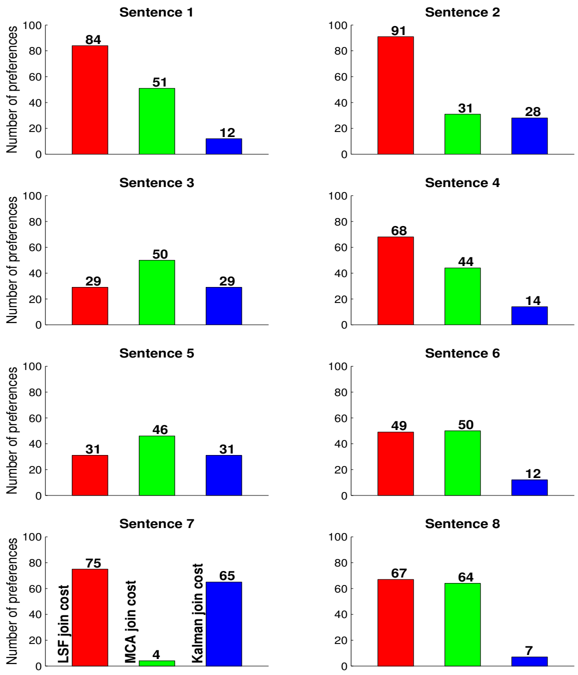

Figure 5: Join cost evaluation, validation cutoff is 10 plus block validation check (after removing invalid blocks)

5.1.1 Paired t-test

We conducted a paired t-test to check the significance of these preference ratings. In this test, preferences for join costs for all sentences (each sentence as a group) were considered. The null hypothesis is that the mean difference $\bar{d}$ between the two join costs is zero; the alternative hypothesis is that it is greater than zero ($\bar{d} > 0$). The test statistic (t) can be computed as follows [25]:

$$t = \frac{\bar{d}}{s/\sqrt{n}} \tag{5}$$

where s is the standard error of the differences and n is the number of groups (in our case $n = 8$). The value of t is compared to the critical values of Students t-distribution with $n - 1$ degrees of freedom to find the probability by chance or significance level (α).

cut-off	*LSF vs MCA*		*MCA vs Kalman*		*LSF vs Kalman*	
	t	α	t	α	t	α
8	1.663	0.20	1.551	0.20	3.831	**0.01**
9	1.591	0.20	1.576	0.20	3.837	**0.01**
10	1.609	0.20	1.401	> 0.2	3.520	**0.01**
11	1.619	0.20	1.465	0.20	3.273	0.02
12	2.161	0.10	2.071	0.10	3.082	0.02
13	0.870	> 0.2	2.296	0.10	2.534	0.05
14	0.764	> 0.2	2.157	0.10	2.454	0.05
15	0.540	> 0.2	0.956	> 0.2	2.308	0.10

Table 5: Paired t-test statistics for the join costs

A two-tailed t-test was used, since we are looking for a preference on either side. In table 5, we present t and α for preference ratings obtained from subjects with validation cutoffs ranging from 8 to 15 (after removing invalid blocks). The preference for LSF join cost over MCA join cost is not statistically significant though the LSF join cost has a greater number of preferences. The preference towards MCA join cost compared to Kalman join cost is also not statistically significant. LSF join cost preferred to Kalman join cost is statistically significant for low validation cutoffs. However, it is less significant for high validation scores (for consistent subject results).

5.1.2 ANOVA results

We also performed a one-way analysis of variance (ANOVA) on preference scores (validation cut-off is 10) of our eight sentences with three levels: LSF join cost, MCA join cost and Kalman join cost. The F value is, $F(2, 21) = 6.77$ which exceeds the critical value, 5.78 (at $\alpha = 0.01$) and p-value < 0.0054. This indicates that there is a significance difference between means of the three join cost functions, i.e. the three join cost functions differ significantly in their listeners' preferences.

In order to determine which pairs of means are significantly different, we conducted a multiple comparison test[7] using MATLAB statistics toolbox. This test revealed that the LSF join cost is significantly ($\alpha = 0.01$) different from Kalman join cost. However, there is no significant difference between LSF join cost and MCA join cost, and between MCA and Kalman join costs.

5.2 Smoothing methods

The preferences for smoothing methods for each sentence are shown in figure 6. Here also we have considered subjects' results, after removing invalid blocks, with validation scores of 10 or more. The preferences for no smoothing and linear smoothing are higher compared to Kalman smoothing. Overall, linear smoothing is preferred more times.

5.2.1 Paired t-test

We present paired t-test statistics for three smoothing comparisons in table 6 for different validation cutoffs (after removing invalid blocks). The preference for no smoothing over linear smoothing is not statistically significant. However there is a significant preference towards linear smoothing over Kalman smoothing except for high validation cutoffs, where it is not significant. Similarly, the preference for no smoothing over Kalman smoothing is significant, but for high validation cutoffs it is less significant.

[7]This test performs a multiple comparison of means or other estimates to determine which estimates are significantly different.

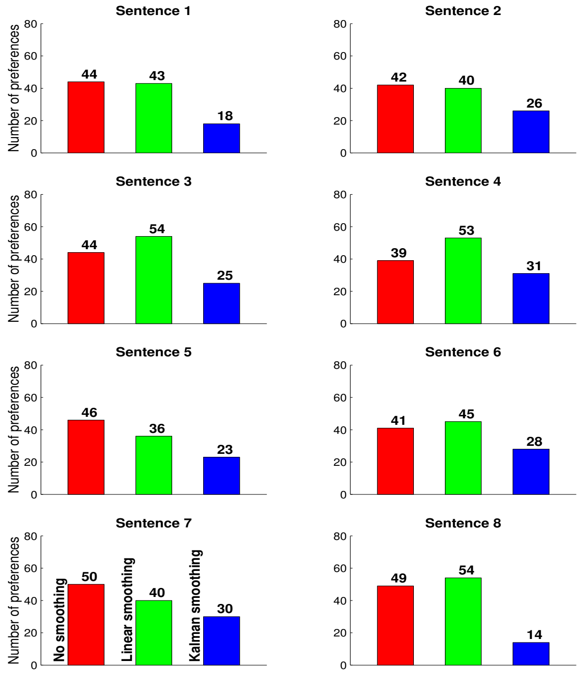

Figure 6: Smoothing evaluation, validation cutoff 10 plus block validation check (after removing invalid blocks)

cut-off	*Linear vs No*		*Linear vs Kalman*		*No vs Kalman*	
	t	α	t	α	t	α
8	1.252	> 0.2	4.330	**0.01**	5.998	**0.01**
9	0.565	> 0.2	4.793	**0.01**	6.450	**0.01**
10	0.406	> 0.2	6.047	**0.01**	6.831	**0.01**
11	0.158	> 0.2	5.133	**0.01**	4.651	**0.01**
12	1.342	> 0.2	2.640	0.05	3.216	0.02
13	0.500	> 0.2	1.730	0.20	2.515	0.05
14	0.205	> 0.2	1.106	> 0.2	1.590	0.20
15	0.607	> 0.2	0.188	> 0.2	0.357	> 0.2

Table 6: Paired t-test statistics for the smoothing methods

5.2.2 ANOVA results

ANOVA (one-way) on preference scores (validation cut-off is 10) of our eight sentences with three levels: no smoothing, linear smoothing and Kalman smoothing resulted in the F value of F(2,21) = 34.05 and the p-value of almost zero. This indicates that the three smoothing methods differ significantly in their listener preferences. We also carried out a multiple comparison test and observed that there is a significant difference between no smoothing and Kalman smoothing and between linear smoothing and Kalman smoothing.

5.3 Kalman-Kalman vs LSF-linear

The preferences for Kalman join cost with Kalman smoothing compared to LSF join cost with linear smoothing are shown in figure 7. LSF-linear is preferred more times than Kalman-Kalman in all sentences. The statistical results in table 7 also conclude that the preference towards LSF-linear is significant.

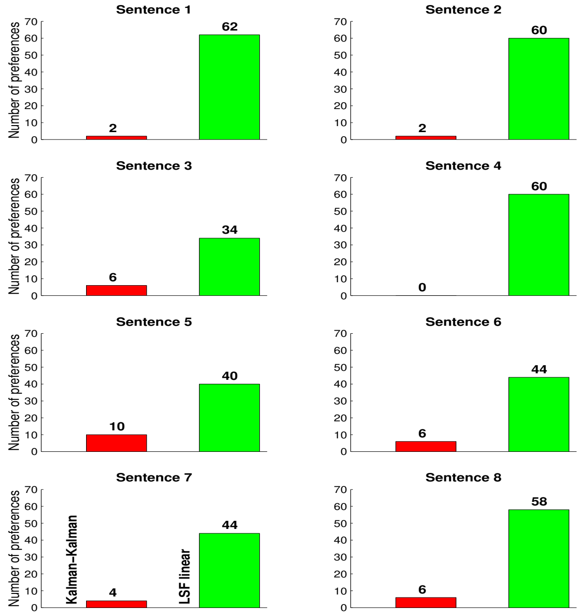

Figure 7: Kalman-Kalman and LSF-linear comparison, validation cutoff 10

cutoff	*LSF-linear vs Kalman-Kalman*	
	t	α
8	8.0958	**0.01**
9	8.7794	**0.01**
10	9.6776	**0.01**
11	8.7767	**0.01**
12	5.9161	**0.01**
13	7.2022	**0.01**
14	3.9886	**0.01**
15	N/A	N/A

Table 7: Paired t-test statistics for the Kalman-Kalman and LSF-linear comparison

6 Conclusions

In this paper, three join cost functions and three different smoothing methods were evaluated by conducting a listening test. In addition to these, combined join cost and smoothing using a Kalman filter was compared with LSF join cost plus linear smoothing.

The results from the listening test indicated that LSF join cost has more preferences than MCA join cost and Kalman join cost. These results re-confirmed our previous perceptual test results (refer table 2). Though the LSF join cost has more preferences, the preference for it over MCA join cost is not statistically significant. The preference towards MCA join cost over Kalman join cost is also not statistically significant. For low validation cutoffs, LSF join cost preference over Kalman join cost is statistically significant. But, for high validation cutoffs (more consistent subjective results) it is less significant.

The rankings of the three join costs in this subjective test are shown in table 8, which agree with the rankings obtained earlier. Therefore we can conclude that the method we proposed in [11, 12, 13] for evaluating join costs based on a single perceptual experiment is further validated.

Rank	*Join Cost*
1	LSF join cost MCA join cost
3	Kalman join cost

Table 8: Rankings for three join costs, obtained in the second listening test

Linear smoothing was preferred more times than no smoothing and Kalman smoothing. There is no significant preference between no smoothing and linear smoothing. However, the preference for both of them over Kalman smoothing is significant except for high validation cutoffs, where the significance is lower. The preference for LSF join cost and linear smoothing over Kalman join cost and Kalman smoothing is statistically significant.

Since the join costs presented here only contain a spectral component, the stimuli presented to listeners in the listening test contained minor F0 discontinuities. It is possible that these discontinuities (partially) mask the effect of spectral discontinuities. This masking provides one possible explanation for cases where listeners had no strong preference, such as between linear smoothing and no smoothing, However, it is simply not known how different factors, such as F0 and spectral envelope, interact in listeners' perception of synthetic speech. This question is the subject of future planned research.

7 Acknowledgements

Thanks to Rhetorical Systems Ltd. for partial funding of this work and the use of *rVoice*. Thanks also to all the experimental subjects: the members of CSTR, Ph.D. students in the Dept. of Linguistics and students on the M.Sc. in Speech and Language Processing, University of Edinburgh.

References

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

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

[3] 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.

[4] G. Coorman, J. Fackrell, P. Rutten, and B. van Coile, "Segment selection in the L & H RealSpeak laboratory TTS system," in *Proc. ICSLP*, Beijing, China, 2000.

[5] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 6, pp. 1983–1986.

[6] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 6, pp. 2747–2750.

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

[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*, Perthshire, Scotland, 2001, pp. 59–62.

[9] A. Crowe and M. A. Jack, "Globally optimising formant tracker using generalised centroids," *Electronic Letters*, vol. 23, no. 19, pp. 1019–1020, 1987.

[10] A. A. Wrench, "Analysis of fricatives using multiple centres of gravity," in *Proc. International Congress of Phonetic Sciences*, 1995, vol. 4, pp. 460–463.

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

[12] 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.

[13] J. Vepa and S. King, "Kalman-filter based join cost for unit-selection speech synthesis," in *Eurospeech*, Geneva, Switzerland, September 2003.

[14] J. Vepa and S. King, "Join cost for unit selection speech synthesis," in *Text to Speech Synthesis: New Paradigms and Advances*, Abeer Alwan and Shri Narayanan, Eds. Prentice Hall, 2004.

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

[16] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[17] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, pp. S35(A), 1975.

[18] F. K. Soong and B. H. Juang, "Line spectrum pairs (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 1.10.1–1.10.4.

[19] Joe Frankel, *Linear dynamic models for automatic speech recognition*, Ph.D. thesis, University of Edinburgh, 2003.

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. Am.Soc.Mech.Eng., Series D, Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[21] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 433–466. Elsevier, Amsterdam, The Netherlands, 1995.

[22] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, The Netherlands, 1997.

[23] David T. Chappell and John H.L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communications*, vol. 36, pp. 343–374, 2002.

[24] A. Black and P. Taylor, "The Festival speech synthesis system: system documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, Univ. of Edinburgh, Edinburgh, Scotland, 1997.

[25] W. John McGhee, *Introductory Statistics*, West Publishing Company, St. Paul, USA, 1985.