# SUBJECTIVE EVALUATION OF JOIN COST & SMOOTHING METHODS

*Jithendra Vepa* and Simon King*

Centre for Speech Technology Research
University of Edinburgh
Edinburgh, UK
www.cstr.ed.ac.uk

## ABSTRACT

In our previous papers, we have proposed join cost functions derived from spectral distances, which have good correlations with perceptual scores obtained for a range of concatenation discontinuities. To further validate their ability to predict concatenation discontinuities, we have chosen the best three spectral distances and evaluated them subjectively in a listening test. The units for synthesis stimuli are obtained from a state-of-the-art unit selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd. We also compared three different smoothing methods in this listening test. In this paper, we report listeners' preferences for each join costs in combination with each smoothing method.

## 1. INTRODUCTION

In unit selection-based concatenative speech synthesis systems, *join cost*, which measures how well two units can be joined together, is one of the main criteria for selecting appropriate units from the large speech database. The perfect join cost should correlate highly with human perception of discontinuity at concatenation boundaries.

In our previous study, we conducted a perceptual experiment to measure this correlation for various join cost functions and reported the results in [1, 2, 3]. 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 use 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]

We start this paper with a description of the join cost functions and smoothing methods used. In section 3, the design and procedure of the listening test is discussed. Finally, we present subjective results of these various combinations and discuss them in section 4.

## 2. JOIN COST FUNCTIONS AND SMOOTHING TECHNIQUES

### 2.1. Join cost functions

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****.*

In previous papers [1, 2, 3] we have presented a method for evaluating join cost function based on the number of statistically significant correlations with perceptual experiment data.

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 1.

*Now at IDIAP, Martigny, Switzerland.

[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.

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

**Table 1**. Rankings for three join costs, obtained in the first listening test

### 2.2. 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.

Linear dynamic models (LDMs)[2], sometimes known as **Kalman filters**, which are used to compute the third of our join cost functions, 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 synthetic waveform. We are investigating 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 compare the Kalman smoothing operation to a linear smoothing technique [5].

#### *2.2.1. Linear smoothing*

The line spectral frequencies (LSF) have good interpolation properties and yield stable filters after interpolation [6]. Although LSF interpolation is widely used in speech coding, it can also be used for speech synthesis. Dutoit [5] showed that LSFs have good interpolation properties and produce smoother transitions than LPC parameters. LSF interpolation was compared with other smoothing methods in [7] 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 [5]. 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 $X_1, X_2, \cdots, X_N$. Assume the number of frames on the left side and the right side of the join to be $M_L$ and $M_R$

[2]LDMs can also be used for speech recognition [4]

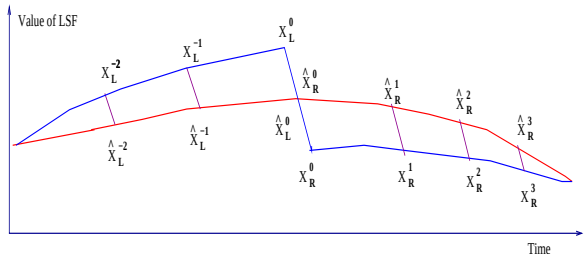


**Fig. 1**. Linear smoothing on parameters (LSFs) of frames at the join (adapted from [5]).

respectively. Then, the LSFs after smoothing ($\hat{X}$) are:

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

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

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.

## 3. 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.*

### 3.1. 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[3], as shown in table 2, which are divided into 12 symmetric[4] blocks.

| | | |
|---|---|---|
| $V_{1a}$-$V_{2a}$<br>$V_{2a}$-$V_{3a}$<br>$V_{3a}$-$V_{1a}$ | $V_{1b}$-$V_{2b}$<br>$V_{2b}$-$V_{3b}$<br>$V_{3b}$-$V_{1b}$ | $V_{1c}$-$V_{2c}$<br>$V_{2c}$-$V_{3c}$<br>$V_{3c}$-$V_{1c}$ |
| $V_{1a}$-$V_{1b}$<br>$V_{1b}$-$V_{1c}$<br>$V_{1c}$-$V_{1a}$ | $V_{2a}$-$V_{2b}$<br>$V_{2b}$-$V_{2c}$<br>$V_{2c}$-$V_{2a}$ | $V_{3a}$-$V_{3b}$<br>$V_{3b}$-$V_{3c}$<br>$V_{3c}$-$V_{3a}$ |
| $V_{1a}$-$V_{2b}$<br>$V_{2b}$-$V_{3c}$<br>$V_{3c}$-$V_{1a}$ | $V_{2a}$-$V_{3b}$<br>$V_{3b}$-$V_{1c}$<br>$V_{1c}$-$V_{2a}$ | $V_{3a}$-$V_{1b}$<br>$V_{1b}$-$V_{2c}$<br>$V_{2c}$-$V_{3a}$ |
| $V_{1a}$-$V_{2c}$<br>$V_{2c}$-$V_{3b}$<br>$V_{3b}$-$V_{1a}$ | $V_{2a}$-$V_{3c}$<br>$V_{3c}$-$V_{1b}$<br>$V_{1b}$-$V_{2a}$ | $V_{3a}$-$V_{1c}$<br>$V_{1c}$-$V_{2b}$<br>$V_{2b}$-$V_{3a}$ |

**Table 2**. All possible pairwise comparisons

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 as 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 the $V_{3c}$. To do this comparison we added the $V_{3c}$ and $V_{1b}$ pair in our test stimuli to the first two rows of table 2.

The test sentences used in our listening test are presented in table 3. These eight sentences were selected randomly from twenty such sentences.

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

[4]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.

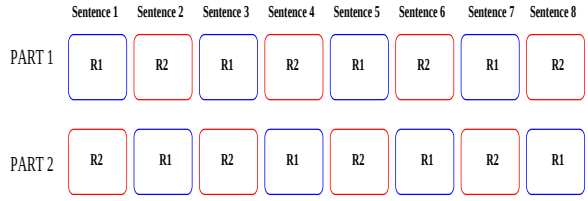


**Fig. 2**. Test procedure, in each part the two rows (R1 and R2) are presented alternatively.

| | |
|---|---|
| *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 3**. Listening test sentences

### 3.2. 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 2, 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 2. In figure 2, 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 around 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 around 15 minutes and about 30-40 minutes for completion

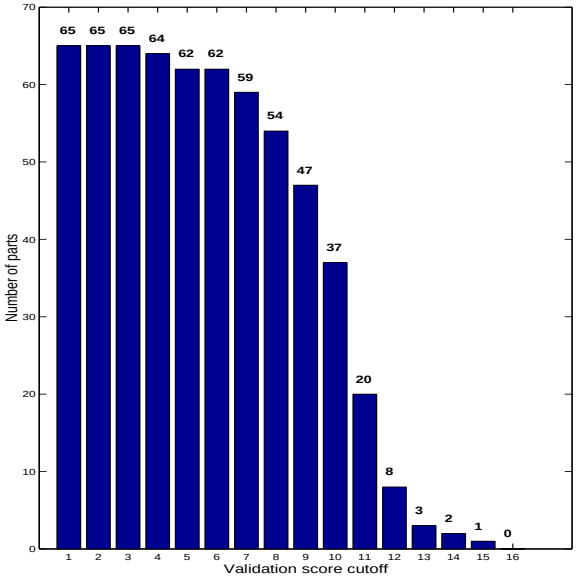


**Fig. 3**. Subjects validity

of two parts. The informal feedback from the subjects indicated that there was not much difference between the two stimuli in many pairs. Infact a few of them felt that those pairs were the same, hence found it a difficult task.

### 3.3. Validation procedures

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 have 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 original and validation pairs. 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 3, 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.

We performed another validation procedure on the block level. Consider the first block in table 2; $V_{1a} - V_{2a}$, $V_{2a} - V_{3a}$ 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.

## 4. SUBJECTIVE EVALUATION

### 4.1. Join costs

In figure 4, 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 least number of preferences.

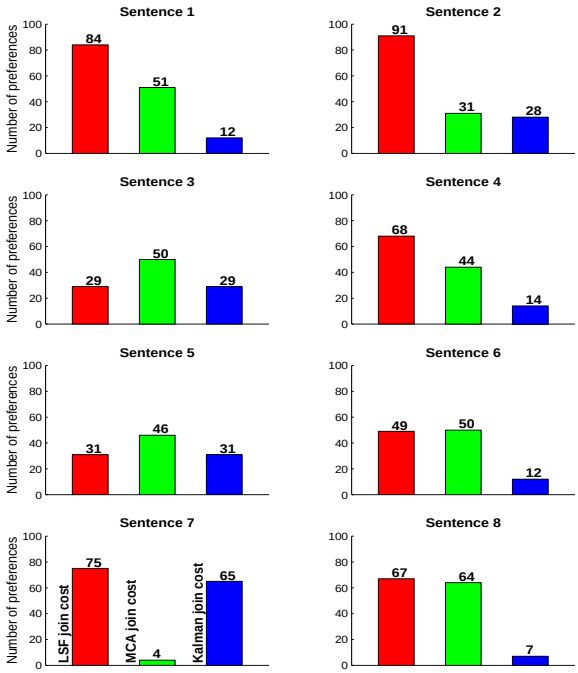


**Fig. 4**. Join cost evaluation, validation cutoff is 10 plus block validation check (after removing invalid blocks)

#### *4.1.1. 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 it is greater than zero ($\bar{d} > 0$). The test statistic ($t$) can be computed as follows [8]:

$$t = \frac{\bar{d}}{s/\sqrt{n}} \tag{3}$$

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 ($\alpha$).

| *cut-off* | *LSF vs MCA* | | *MCA vs Kalman* | | *LSF vs Kalman* | |
|---|---|---|---|---|---|---|
| | $t$ | $\alpha$ | $t$ | $\alpha$ | $t$ | $\alpha$ |
| 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 4**. 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 4, we present $t$ and $\alpha$ 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).

### 4.2. Smoothing methods

The preferences for smoothing methods for each sentence are shown in figure 5. 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.

We present paired t-test statistics for three smoothing comparisons in table 5 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.

### 4.3. 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 6. LSF-linear is preferred more times than Kalman-Kalman in all sentences. The statistical results in table 6 also conclude that the preference towards LSF-linear is significant.

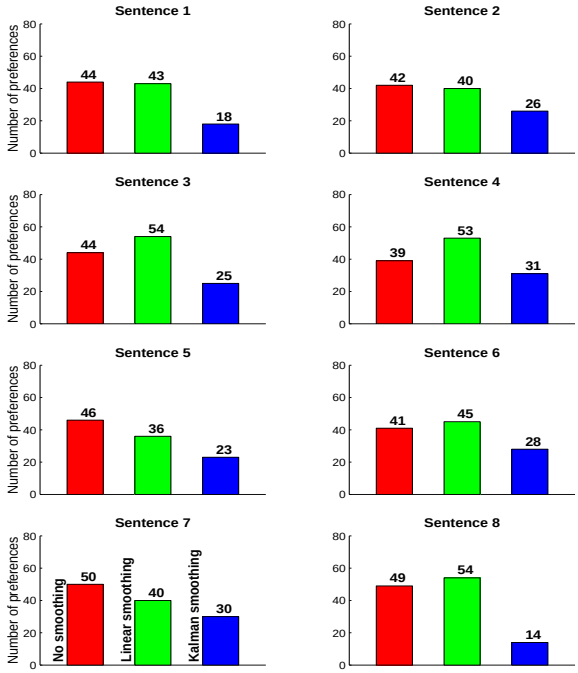


**Fig. 5**. Smoothing evaluation, validation cutoff 10 plus block validation check (after removing invalid blocks)

| *cut-off* | *Linear vs No* | | *Linear vs Kalman* | | *No vs Kalman* | |
|---|---|---|---|---|---|---|
| | $t$ | $\alpha$ | $t$ | $\alpha$ | $t$ | $\alpha$ |
| 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 5**. Paired t-test statistics for the smoothing methods

## 5. 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 reconfirmed our previous perceptual test results (refer table 1). 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 statisti-

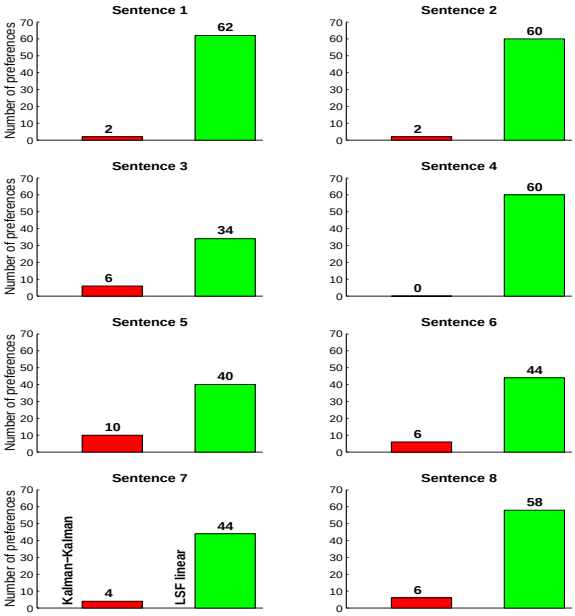


**Fig. 6**. Kalman-Kalman and LSF-linear comparison, validation cutoff 10

| *cutoff* | *LSF-linear vs Kalman-Kalman* | |
|---|---|---|
| | $t$ | $\alpha$ |
| 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 6**. Paired t-test statistics for the Kalman-Kalman and LSF-linear comparison

cally 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.

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.

The rankings of the three join costs in this subjective test are shown in table 7, which agrees with the rankings obtained earlier. Therefore we can conclude that the method we proposed in [1, 2, 3] for evaluating join costs based on a single perceptual experiment is successful.

| *Rank* | *Join Cost* |
|---|---|
| 1 | LSF join cost<br>MCA join cost |
| 3 | Kalman join cost |

**Table 7**. Rankings for three join costs, obtained in the second listening test

## 6. 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.

## 7. REFERENCES

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

[2] 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.

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

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

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

[6] 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.

[7] 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.

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