IDIAP RESEARCH REPORT

Subjective Evaluation of Join Cost Functions Used in Unit Selection Speech Synthesis

Jithendra Vepa [a] Simon King [b]

IDIAP–RR 04-26

September 2004

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Research Institute, CH-1920, Martigny, Switzerland.
[b] Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK.

IDIAP Research Report 04-26

Subjective Evaluation of Join Cost Functions Used in Unit Selection Speech Synthesis

Jithendra Vepa Simon King

September 2004

submitted for publication

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 unit sequences 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 of the three join cost functions.

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 [1, 2, 3]. The perfect join cost should correlate highly with human perception of discontinuity at unit 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 [4, 5, 6].

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. We used our own implementation of residual excited linear prediction (RELP) synthesis for waveform generation using the unit sequence selected by the experimental version of *rVoice* synthesis system.

We start this paper with a description of the join cost functions evaluated subjectively. Also, we explain the implementation of the RELP resynthesis method. In section 3, the design and procedure of the listening test is discussed. Finally, we present subjective results of these join cost functions and discuss them in section 4.

2 Join cost functions & Waveform generation

2.1 Join cost functions

We have chosen three of the best spectral distances, which were used in the join cost functions, from our previous papers [4, 6] based on the number of statistically significant correlations with perceptual experiment data. 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 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 listed above scored **six** 1% significant correlations out of a possible maximum of ten. 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.

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

Table 1: Rankings for three join costs, obtained in our previous perceptual tests

2.2 Residual excited linear prediction

Residual excited LPC (RELP) is one of the standard methods for resynthesis, which is also used in Festival [7]. In this method, first LPC analysis has to be carried out on the original speech to obtain LPC parameters. During LPC analysis we have computed the LPC parameters using asymmetric[1] Hanning-windowed pitch-synchronous frames of the original speech as shown in figure 1. 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 1.

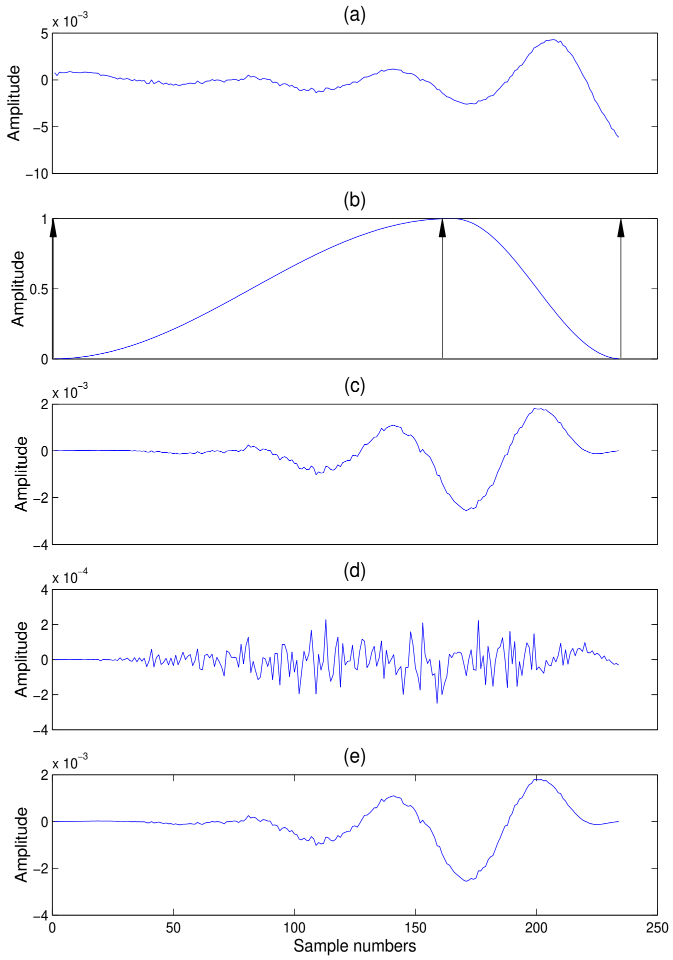

Figure 1: RELP resynthesis using an asymmetric window: (a) Original waveform (b) Asymmetric Hanning window (pitch marks shown as arrows) (c) Windowed original waveform (d) Residual signal (e) Reconstructed waveform

Once the units are selected using the *rVoice* synthesis system, the corresponding LPCs and residual signals from the database are assembled. Then, the LPC filter is excited using the residual to reconstruct the output speech waveform. In figure 1, the output waveform is depicted in the last plot, which is a near-perfect reconstruction of the original signal. To get the full synthetic waveform for an utterance we overlap and add these two-pitch-period waveforms.

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

3 Listening test

A listening test was designed to evaluate the three join cost functions: LSF join cost, MCA join cost and Kalman join cost. To know which join cost performs better, we need to do three pair-wise comparisons, which are:

1. LSF join cost (V_1) *vs* MCA join cost (V_2)
2. MCA join cost (V_2) *vs* Kalman join cost (V_3)
3. Kalman join cost (V_3) *vs* LSF join cost (V_1)

where V_1, V_2 and V_3 are synthesised versions using three join cost functions: LSF, MCA and Kalman join costs respectively.

3.1 Test stimuli

The test sentences used in our listening test are presented in table 2. 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 2: Listening test sentences

3.2 Test procedure

There were 33 participants in this listening test. Most of them were members of 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. On the average, subjects took around 15 minutes for completion. 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.

3.3 Validation procedures

To check the validity of the subjects' results, we included 16 validation pairs[2] in 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 2, 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.

[2]Each pair means one comparison, for example $V_1 - V_2$

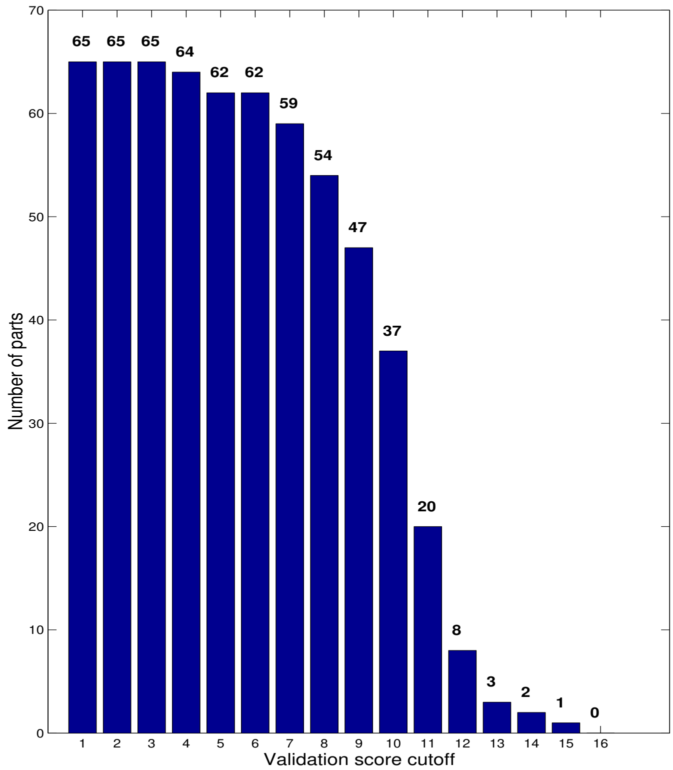

Figure 2: Subjects validity

We performed another validation procedure on the block level. Consider the block as; $V_1 - V_2$, $V_2 - V_3$ and $V_3 - V_1$. If subjects preferred all the first stimuli (V_1, V_2 and V_3) then the block becomes invalid because, if they prefer V_1 and V_2, then for the third pair, the valid selection is V_1. Similarly, they can not prefer all the second stimuli in a block.

4 Subjective evaluation

In figure 3, 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.

4.1 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 ($\bar{d} \neq 0$). The test statistic (t) can be computed as follows [8]:

$$t = \frac{\bar{d}}{s/\sqrt{n}} \tag{1}$$

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

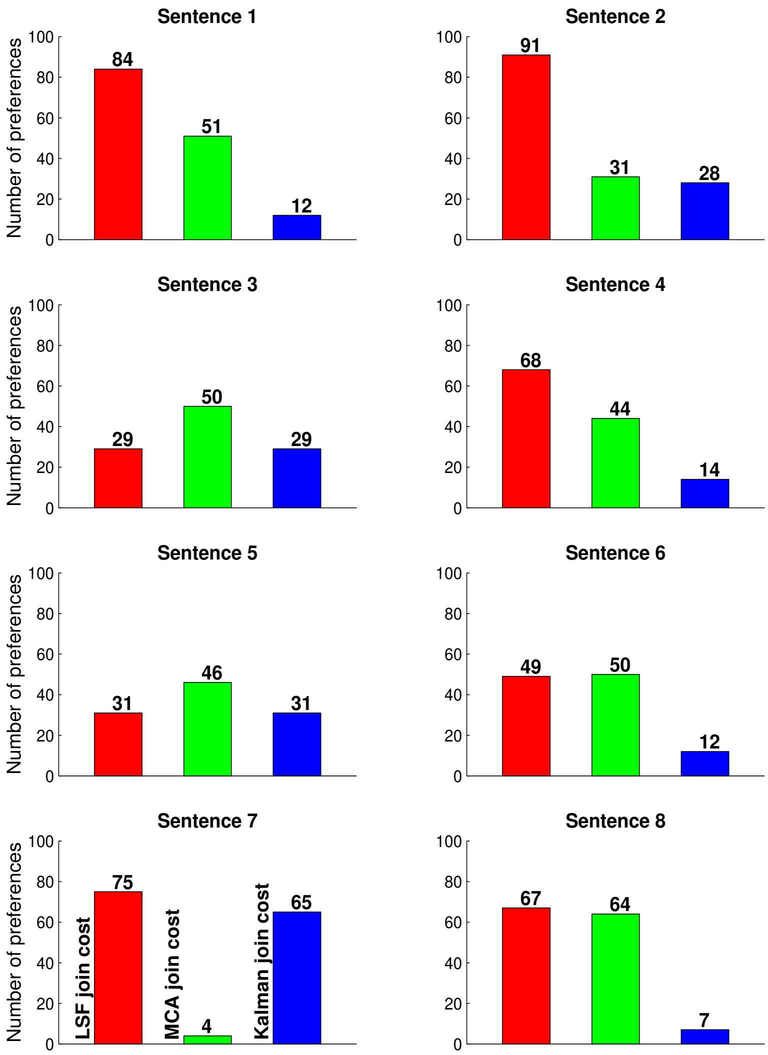

Figure 3: Join cost evaluation, validation cutoff is 10 plus block validation check (after removing invalid blocks)

to find the probability by chance or significance level (α). Low probability values ($\alpha \leq 0.01$) reject the null hypothesis and one can say the preference for a particular join cost is statistically significant.

A two-tailed t-test was used, since we are looking for a preference on either side. In table 3, 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).

4.2 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.

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 3: Paired t-test statistics for the join costs

The F value is, $F(2, 21) = 6.77$ which exceeds the critical value, 5.78 (at $\alpha = 0.01$) and $p < 0.0054$. This indicates that there is a significance difference between means of the three join cost functions, i.e. 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 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 Conclusions

In this paper, three join cost functions were evaluated by conducting a listening test. 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 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 4, which agrees with the rankings obtained earlier. Therefore we can conclude that the method we proposed in [4, 5, 6] for evaluating join costs based on a single perceptual experiment is successful.

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

Table 4: Rankings for three join costs, obtained in the current 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.

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] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[3] 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.

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

[5] 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.

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

[7] 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.

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