Investigating Festival's target cost function using perceptual experiments

Volker Strom, Simon King

Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

vstrom@inf.ed.ac.uk

Abstract

We describe an investigation of the target cost used in the Festival unit selection speech synthesis system [1]. Our ultimate goal is to automatically learn a perceptually optimal target cost function. In this study, we investigated the behaviour of the target cost for one segment type. The target cost is based on counting the mismatches in several context features. A carrier sentence ("My name is Roger") was synthesised using all 147,820 possible combinations of the diphones /n_ei/ and /ei_m/. 92 representative versions were selected and presented to listeners as 460 pairwise comparisons. The listeners' preference votes were used to analyse the behaviour of the target cost, with respect to the values of its component linguistic context features.

Index Terms: speech synthesis, unit selection, target costs

1. Introduction

The speech database for a unit selection voice typically consists of several hours of speech from a single speaker. A text selection algorithm is used to select the text to record. This algorithm is typically hand-crafted and aims for wide coverage of all units in as many different contexts as possible. The context of a unit is defined by a set of features such as the preceding and following unit identities, phrase position, stress, etc. Ideally the database would contain multiple examples of each unit in all contexts, in order to find perceptually acceptable concatenations — this is, of course, an impossible goal.

During synthesis, a *target cost function* is used to select units from the database. Again, this is based on matching context features of the units, typically as a weighted sum of penalty terms. Target cost functions are manually-designed and difficult to optimise. Tuning the many weights of this highly non-linear function (which may have multiple local optima) by hand is unlikely to produce optimal results. Furthermore, these conventional cost functions are usually *linear* combinations of factors. A linear relationship between the context features and listeners' perceptual ratings seems highly unlikely!

Learning the weights of a typical target cost function from data is also very challenging. Optimising the weights typically involves iterative weight updates based on subjective data (listening judgements) and/or 'objective' measurements (comparison with natural utterances). The use of natural target utterances (typically with a spectral distance measure) fails to acknowledge that there is always more than one acceptable way to render a given utterance. Iterated weight updates and listening tests is very time consuming and may be ill-posed: there is no proof that a single setting of the weights can work in all contexts.

In contrast to this approach, our goal is to learn a target cost function that, rather than placing a fine-grained continuously-valued cost on each candidate unit, makes *perceptual acceptability* judgements. These might be much more coarse-grained, for example on a 3 point scale ("very good", "adequate", "unacceptable"). The target cost function would then be a *classifier* and not a continuous function. This greatly widens the possibilities for types of model that can be considered, may simplify learning the model from data, and may require less data for learning. The key challenge is to establish the relationship between the linguistic context features and perceptual acceptability. Databases used in speech synthesis will always have missing units (in terms of context feature combinations), but a missing unit is not a problem if a perceptually equivalent unit exists elsewhere in the database, and we know how to select it based on its linguistic features. Once this is possible, the text-selection algorithm, target cost and back-off strategy can exploit this knowledge in a consistent way.

left_context	each phone or none
right_context	each phone or none
position_in_word	initial, mid, final, inter
position_ in_syllable	initial, mid, final, inter
part_of_speech	nouns, verbs, function words, adjectives/adverbs/particles
position_in_phrase	within phrase-final syllable, or not
stress	none, primary, secondary, ternary
boundary	none, continuation, terminal, interrogative
emphasis	yes, no

Table 1: *The linguistic context features used in our target cost function, with their possible values. All features are associated with phone units, not diphones, except for position_in_word and position_in_syllable (where "inter" means the diphone crosses a boundary). The target cost function compares the feature values for both constituent demiphones of a candidate diphone with those of the target diphone. Stress, boundary and emphasis take a default value if the phone is not a vowel.*

We plan to create a set of listeners' perceptual judgements for a particular voice, from which perceptual acceptability under various combinations of context features can be learnt without performing further listening tests. In the pilot study reported here, we focus on just one phone, the diphthong /ei/ (as in "name"; the phone set is that of the RP variant of our Unisyn [3] lexicon).

2. Significance Tests

This initial phase of the experiment is intended to simplify the subsequent perceptual evaluation: if changing the value of a context feature does not cause any *acoustic* difference, there will not be *perceptual* difference either. The features used in our target cost function, and their possible values, are shown in Table 1. We investigated whether instances of /ei/ having a particular context feature value are *acoustically significantly dif-*

ferent from other instances. We parameterised the speech using 14 parameters: 12 MFCCs, F0 and log energy. Each coefficient is normalised such that their mean is 0.5, $-4stddev$ is 0 and $+4stddev$ is 1. We computed the distribution of each acoustic parameter at the mid-point of /ei/ and used the Mann-Whitney U test to look for significant differences in the distribution of the acoustic parameter.

Initially, we compared the distribution for instances of /ei/ with a particular feature value to the distribution for all instances (i.e., the global distribution). As might be expected, the sample sizes tend to be very uneven (either very small or very large); e.g., almost all instances of /ei/ have primary stress, so it is not surprising that there is no significant difference in any of the 14 acoustic parameters between all instances of /ei/ and those instances with primary stress.

Therefore, we chose to compare possible feature value pairs within each feature (rather than comparing a single feature value to the global distribution), which results in far more comparisons (particularly for phonemic context). The results revealed a significant difference in the acoustic parameters between instances with one value for a particular feature and those with another value for that feature in almost all cases. But this does not imply that they are perceptually different: the sample sizes are very large, and the U test is sensitive to even small difference in the distributions. However, a few entirely redundant distinctions were discovered: e.g., ternary stress and no stress are indistinguishable for all 14 acoustic parameters at a 99% confidence level, and could therefore safely be collapsed into a single class. We are assuming that our 14 acoustic parameters capture all perceptually salient information in the speech signal.

The phonetic context components of our target cost test only for an exact match; all mismatches are considered of equal importance and incur the same penalty which is added into the total target cost. For /ei_m/, each of the 14 acoustic parameter distributions for instances with a left_context feature value of /n/ are indistinguishable from the distributions of those instances with a left_context feature value /ou/, /@@r/, /e/ and /uh/. If we require only 13 of the 14 acoustic parameters to be indistinguishable, /n/ also becomes indistinguishable from /@@/, /ei/, /ii/ and /ai/. right_context behaves similarly. Pairwise statistics do not imply mutual equivalence between /ou/, /@@r/, /e/ and /uh/ (for example), so we cannot group them into a class.

3. Experiment

In the next part of our experiment, we focus again on the diphthong /ei/, this time in a particular context. Following [4], we varied the diphones chosen from the database to render /n_ei/ and /ei_m/ (which constitute the diphthong /ei/) in "My name is Roger". The other diphones were fixed; there are thus 3 concatenation points and 4 demiphones that vary; everything else is constant. The goal is to obtain perceptual ratings for many different combinations of context feature values, to discover the relationship between features and perceptual acceptability.

3.1. Selection of test materials

There are 389 candidates in the database for /n_ei/ and 380 for /ei_m/, which results in 147,820 possible versions of the carrier sentence. It is clearly not possible to present all of these to listeners, so we preselected those versions with low join costs at all three of the concatenation points that could differ between versions. Ideally, we would like a set of versions of the carrier sentences with imperceptible joins, so that we could be sure that listeners' ratings were affected only by the context features. Thresholding the join cost narrowed the number of versions down to 124. We then reduced this to 92 versions, by eliminating versions with duplicate patterns of feature mismatches.

cumulative votes		Mismatching feature
20.9%	18 of 86	emphasis_2
34.0%	284 of 836	emphasis_1
39.4%	614 of 1557	position_in_syllable_2
39.4%	614 of 1557	position_in_word_2
40.7%	474 of 1164	boundary_1
48.7%	2683 of 5509	position_in_phrase_1
48.8%	3421 of 7014	left_context_2
49.3%	768 of 1558	stress_1
49.7%	2921 of 5881	partofspeech_2
49.8%	4053 of 8144	position_in_syllable_1
49.8%	1815 of 3642	position_in_word_1
50.5%	4362 of 8632	right_context_1
50.6%	1458 of 2882	position_in_phrase_2
51.5%	1642 of 3189	partofspeech_1
56.8%	413 of 727	stress_2

Table 2: *Feature ranking, based on cumulative votes, in percent and absolute numbers, for a mismatch in a particular feature. Lower numbers indicate higher importance.*

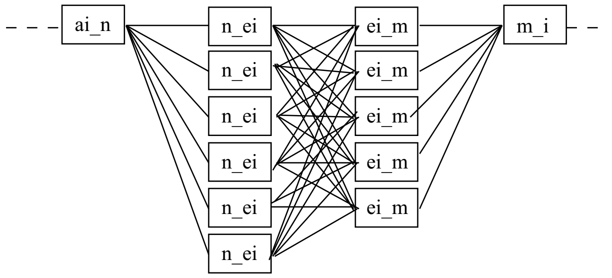

Figure 1: *Varying the two diphones which constitute the diphthong /ei/ in "My name is Roger" to create many versions of the same sentence.*

462 pairs were selected from the 4186 possible pairs such that listener ratings of these pairs will be most informative with respect to a particular context feature: ideally, the two versions constituting a pair differed only in one feature's value, with one of them matching the target specification and one not matching, and with all remaining features matching the target specification exactly. Where that was not possible, then the remaining features had to either both match or both mismatch the target specification. Failing that, the number of differences (i.e., where a feature of one version matched the target and the other did not) had to be as small as possible.

For simplicity, only the features of the two /ei/ demiphones of /ei/ were considered during this selection procedure. Our full target cost would normally also consider the features of the /n/ demiphone in the left diphone and the /m/ demiphone in the right diphone. Note that our synthesiser joins diphthongs not in the middle but at the 25% point.

3.2. Test procedure

We conducted a pairwise forced choice preference test, rather than asking listeners to provide opinion score ratings (e.g., on a

5 point scale) for individual versions, because we felt this was a simpler task for listeners and was more likely to produce consistent judgements. 43 listeners were used; all were native English speakers. 120 pairs were presented to each listener over headphones using via a web browser in a quiet environment supervised by the experimenter. Listeners could listen to each version repeatedly if they wished.

3.3. Results

The results of the listening test were summarised as the number of votes that each of the 92 versions received – that is, the number of times that version was preferred over some other version. We confirmed that the maximum join cost (that is, the highest value amongst the costs of the three variable concatenation points) was not correlated with the number of votes (r=-0.008). In the following discussion, suffixing a feature name with _1 or _2 means that feature is for the first or second demiphone of /ei/, respectively.

From the vote counts, we calculated the percentage of times that versions with a mismatched feature (i.e., one that does not match the target specification) were actually preferred by listeners. If a feature is perceptually important, then this percentage should be low. For example, if listeners never like versions where feature emphasis_2 mismatches the target specification, then such versions will receive no votes.

Starting with the feature that has the lowest such percentage (i.e., the feature of a candidate that is most the important to match with the target) and working upwards, Table 2 lists the features from the most important to the least important. Only 15 of the 18 features are shown because left_context_1 and right_context_2 were disregarded along with all other features of the varied demiphones in /n/ and /m/, and because there were no mismatching boundary_2 features in the sentences used in the listening test.

The table shows that a mismatch in the emphasis feature (of either left or right demiphone) is very unpopular. Also, the features of the second demiphone seem to be generally more important than those of the first demiphone. A problem with this approach is that the cumulative votes for each feature are pooled over all other features (regardless of whether they match or mismatch). Also, the method is informative about the few most important features (emphasis_2, emphasis_1, ...), but is less helpful in distinguishing amongst those of lesser importance (e.g., stress_2)

To evaluate a feature, we would ideally use listener preferences of only those pairs which differ in a match of that feature. For 4 of our 18 features, there is no such pair, and for others there are only a few. So we must pool across multiple pairs – we have to allow a few other features to vary in their match/mismatch status. We do this stepwise, beginning with the least important other feature (working upwards in Table 2).

Table 3 shows the results (position_in_phrase_2 is omitted to save space; the results vary from 42.9% of 212 to 48.1% of 833). Generally one would expect the percentage to converge towards 50% moving down this table, as results are pooled over more and more features. When this does not occur, it indicates that the feature under investigation (column header) is perceptually important, whatever is happening with other features. For each feature and pooling, we tested if F0 discontinuities in /ei/ or the number of mismatches in other features coincidentally favoured one version, but did not find any such effect.

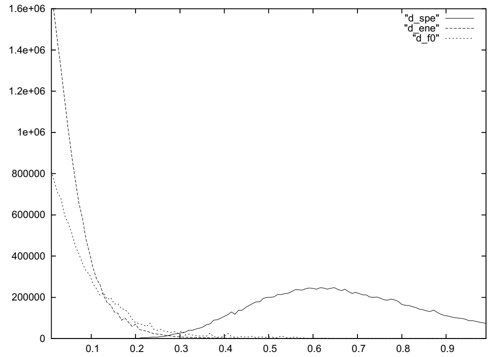

Figure 2: *Distribution of the 3 join cost components: F0 distance, energy distance and spectral distance (mean distance over 12 MFCCs), with distances 0 and 1 masked out.*

3.4. Looking at the join cost again

Although we had previously confirmed that the accumulated votes for a version did not correlate with the version's maximum join cost, we suspected that there might still be correlations with individual components of the join cost. In an experiment following [2], we synthesised 450,000 newspaper sentences and obtained statistics of target and join costs and their components. Although each coefficient had been normalised, figure 2 shows that the distributions of the three join cost components are not comparable numerically: the spectral distance is on average 7.7 greater than the F0 distance (the special distances 0 and 1 are masked out; 1 as F0 distance indicates that one demiphone is voiced and the other is not).

However, even if the join cost components were normalised w.r.t. their means and standard deviations, they would still not be comparable perceptually. Table 4 shows, for each of the three varying concatenation points, the correlation between the accumulated votes and the join cost components, their weighted sum (which is currently the average), and, in the rightmost column, the correlation with the maximum join cost over the three concatenation points.

It appears that preselecting the versions for use in the listening test by thresholding the maximum join cost was inappropriate, because that measure has the least correlation with listeners' votes. The join in /ei/ is more important than the other joins; in particular, F0 join cost correlates with listeners' votes most (negatively – lower join cost is preferred). We conclude that

	/n/	/ei/	/m/	max
d_spe	0.059	0.153	-0.061	-0.039
d_ene	-0.117	-0.032	0.197	0.025
d_f0	0.091	**-0.339**	-0.133	-0.255
avrg	0.050	-0.015	-0.046	**-0.008**

Table 4: *The correlation of the accumulated votes with the spectral, energy and F0 components of the join costs as well as the total join cost (the average of those three components) at each of the three varying concatenation points. The lower rightmost cell is the correlation mentioned earlier. Values highlighted in bold face are discussed in the text.*

	boundary_1	emphasis_1	emphasis_2	partofspeech_1	partofspeech_2	left_context_2	pos_in_phrase_1
no pooling	72% of 79	74% of 54	no votes	56% of 155	46% of 78	56% of 187	51% of 286
stress_2	72% of 79	74% of 54	no votes	56% of 155	50% of 185	54% of 266	51% of 298
partofspeech_1	72% of 79	71% of 89	no votes	n/a	50% of 185	52% of 430	52% of 377
pos_in_phrase_2	63% of 133	64% of 174	no votes	57% of 167	53% of 384	54% of 507	52% of 377
right_context_1	63% of 133	64% of 174	no votes	56% of 178	54% of 395	54% of 575	52% of 377
pos_in_word_1	63% of 201	66% of 236	no votes	47% of 320	54% of 437	55% of 607	55% of 442
pos_in_syl_1	66% of 232	66% of 236	no votes	50% of 398	54% of 448	55% of 607	55% of 453
partofspeech_2	65% of 296	68% of 268	no votes	50% of 398	n/a	56% of 661	56% of 517
stress_1	66% of 317	68% of 279	no votes	50% of 498	54% of 459	55% of 683	56% of 598
left_context_2	67% of 360	74% of 432	no votes	52% of 564	55% of 623	n/a	56% of 689
pos_in_phrase_1	68% of 562	74% of 499	no votes	51% of 628	54% of 702	55% of 774	n/a
boundary_1	n/a	73% of 575	62% of 43	48% of 724	53% of 745	56% of 817	59% of 891
pos_in_word_2	68% of 562	73% of 575	62% of 43	48% of 724	53% of 745	56% of 817	59% of 891
pos_in_syl_2	68% of 562	72% of 598	79% of 86	47% of 789	55% of 766	53% of 869	56% of 1000
emphasis_1	65% of 638	n/a	79% of 86	44% of 865	52% of 874	58% of 1022	56% of 1056
emphasis_2	65% of 692	72% of 598	n/a	44% of 865	51% of 917	58% of 1022	56% of 1056

	pos_in_syl_1	pos_in_syl_2	pos_in_word_1	pos_in_word_2	right_context_1	stress_1	stress_2
no pooling	no votes	no votes	35% of 45	no votes	57% of 66	40% of 87	27% of 54
stress_2	no votes	no votes	35% of 45	no votes	57% of 66	40% of 87	n/a
partofspeech_1	81% of 11	no votes	46% of 87	no votes	50% of 77	40% of 87	32% of 161
pos_in_phrase_2	81% of 11	no votes	51% of 119	no votes	50% of 77	42% of 119	35% of 183
right_context_1	46% of 213	no votes	51% of 119	no votes	n/a	41% of 221	34% of 194
pos_in_word_1	46% of 224	no votes	n/a	no votes	50% of 77	50% of 308	36% of 226
pos_in_syl_1	n/a	no votes	52% of 130	no votes	46% of 290	48% of 428	38% of 237
partofspeech_2	49% of 312	no votes	45% of 284	no votes	47% of 301	48% of 528	39% of 259
stress_1	51% of 442	no votes	45% of 415	no votes	45% of 403	n/a	39% of 259
left_context_2	52% of 463	no votes	47% of 479	no votes	45% of 471	48% of 550	39% of 414
pos_in_phrase_1	53% of 530	no votes	49% of 579	no votes	45% of 471	48% of 631	42% of 516
boundary_1	50% of 561	no votes	51% of 647	no votes	42% of 524	49% of 652	41% of 536
pos_in_word_2	50% of 561	79% of 525	51% of 647	n/a	42% of 524	49% of 652	41% of 536
pos_in_syl_2	50% of 561	n/a	51% of 647	77% of 470	40% of 556	49% of 652	42% of 557
emphasis_1	51% of 572	77% of 548	54% of 744	77% of 470	40% of 556	48% of 675	41% of 578
emphasis_2	51% of 572	78% of 581	54% of 744	78% of 491	41% of 566	51% of 751	41% of 578

Table 3: *Cumulative votes for feature matches when pooling over more and more features, beginning with the least important.*

combining join cost components by averaging is a poor match to listeners' perceptions.

4. Discussion

The most striking result was that most context features seem to play a minor role (e.g. left_context_2) or none at all (stress). Note that the POS tagger in Festival is no longer state-of-the-art, which may explain why the POS features are not useful. We constructed a simplified version of the target cost function, in which only the emphasis and the boundary tone features were considered. In informal listening, the resulting synthetic speech appeared generally more natural in terms of the contextual appropriateness of the units and certainly in terms of continuity of F0 across concatenation points. The latter is not surprising, because a longer list of candidates is effectively available for each target unit, allowing the join cost to choose those with better joins. However, it appears that the unit selections become more sensitive to bad labelling in the database. We speculate that the components of the target cost that we removed were somehow mitigating this bad labelling, rather than directly selecting the most appropriate units. The interaction of the target cost and errors in the database labelling was not considered in our experimental design, but deserves further investigation.

The other important result is that our join cost should take the maximum value over its component sub-costs, rather than a weighted sum. This is intuitively reasonable, because human perception does not operate as a (linear) weighted sum; rather, a small number of cues tend to dominate listeners' judgements.

Phonemic context plays a small but significant role, but would no doubt be more useful if the target cost was expressed in terms of appropriate phonetic *classes* instead. Surprisingly, stress is not significant (the acoustic differences are statistically significant but small in magnitude).

5. Acknowledgements

Volker Strom is funded by EPSRC grant EP/E031447/1. Simon King holds an EPSRC Advanced Research Fellowship.

6. References

[1] Clark R., Richmond K. and King S.,"Multisyn voices from ARCTIC data for the Blizzard challenge", Proc. Interspeech, 2007.

[2] Beutnagel M. and Conkie A., "Interaction of Units in a Unit Selection Database", Proc. Eurospeech, 1999.

[3] Fitt S. and Isard S., "Synthesis of regional English using a keyword lexicon", Proc. Eurospeech 1999.

[4] Vepa J., King S. and Taylor P., "Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis", Proc. ICSLP 2002.