SMOOTH TALKING: ARTICULATORY JOIN COSTS FOR UNIT SELECTION

Korin Richmond and Simon King

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

ABSTRACT

Join cost calculation has so far dealt exclusively with acoustic speech parameters, and a large number of distance metrics have previously been tested in conjunction with a wide variety of acoustic parameterisations. In contrast, we propose here to calculate distance in articulatory space. The motivation for this is simple: physical constraints mean a human talker's mouth cannot "jump" from one configuration to a different one, so smooth evolution of articulator positions would also seem desirable for a good candidate unit sequence. To test this, we built `Festival Multisyn` voices using a large articulatory-acoustic dataset. We first synthesised 460 TIMIT sentences and confirmed our articulatory join cost gives appreciably different unit sequences compared to the standard `Multisyn` acoustic join cost. A listening test (3 sets of 25 sentence pairs, 30 listeners) then showed our articulatory cost is preferred at a rate of 58% compared to the standard `Multisyn` acoustic join cost.

Index Terms— speech synthesis, unit selection, electromagnetic articulography, join cost

1. INTRODUCTION

Over the past ten years or so, attention within the speech synthesis research community has become increasingly focussed upon statistical parametric acoustic modelling methods, with work first based mainly on the hidden Markov model (HMM) [1] or Classification and Regression Tree clustering [2] for example, but more recently on a widening array of machine learning models. Deep neural networks (DNNs) (e.g. [3, 4, 5, 6]) have in particular become very popular, but other models such as random forests [7] or linear dynamic models [8] have also appeared. This is understandable, as statistical parametric methods promise robust and fully flexible speech synthesis with a voice building process that may be largely automated. Statistical parametric speech synthesis (SPSS) methods have indeed already demonstrated impressive performance in terms of intelligibility and consistency in numerous studies. At the same time, however, it is recognised that the quality of statistical parametric voices requires further improvement in terms of naturalness. Zen et al. [1], for example, pointed to three major factors that degrade the quality of the speech synthesised by SPSS methods: vocoding, accuracy of acoustic models, and over-smoothing.

In contrast to SPSS methods, speech from the leading alternative synthesis method, unit selection waveform concatenation [9], can demonstrate a very high degree of naturalness. Unit selection aims to "join" pre-recorded fragments of human speech back together with minimal signal processing. To build a unit selection voice, a single speaker is first recorded uttering a large number of phonetically diverse sentences. These utterances are then chopped up into "units", with accurate labelling in terms of phone timings and linguistic structure, to give the voice unit database. To synthesise a new utterance, the unit database is searched to find the best sequence of candidate units to match the desired target sequence. Errors may sometimes occur in selecting and joining units, which lead to spurious glitches, but in the majority of cases the process can work well. We therefore find that, with sufficient care in constructing and using the unit database, the synthetic speech can sound extremely similar to the original human speaker. For these reasons, and especially where CPU and memory constraints are not a significant concern, unit selection is still prevalent in most high quality commercial systems, and thus remains a very relevant synthesis technique.

Searching the unit selection voice database for the ideal candidate unit sequence is typically done with a variant of a Viterbi search, guided by two cost functions: the target and join costs respectively. The join cost is responsible for predicting how well two pre-recorded units of speech will join together in sequence, seeking to avoid any perceptible discontinuities. All of the previous work of which we are aware has concentrated on calculating join cost functions in the acoustic domain. These join costs typically have subcomponents to measure three types of (mis-)match across a potential join point: f0, energy/loudness, and spectral match.

Previous work to improve join costs has largely taken f0 and energy matching as fixed, and looked primarily at spectral matching. We can discern three main themes in this work: i) trying different acoustic parameterisations; ii) trying different distance metrics; iii) trying to tailor acoustic join cost calculation to maximally correlate with human perception of the obtrusiveness of unit joins (e.g. [10, 11, 12, 13]). Stylianou and Syrdal [12] started with a perceptual test using human listeners to derive join discontinuity detection rates for a range of synthetic stimuli. They then evaluated 13 different objective distance measures with respect to their ability to predict the detection rates derived from the human listener ratings.

Corresponding author email: `korin@cstr.ed.ac.uk`. This work was supported by EPSRC grants EP/I031022/1 (Natural Speech Technology) and GR/R94688/01 (Cougar).

The criteria they used to classify these distances included detection rate, the Bhattacharyya measure of distribution separability, and receiver operating characteristic curves. In their results, Kullback-Leibler distance between power spectra gave the highest detection rate, followed by Euclidean distance calculated for mel-frequency cepstral coefficients (MFCCs). Vepa and King [13] also used a similar approach of collecting human judgements of join discontinuities to correlate objective join cost scores with, but also looked at join smoothing too. They used MFCCs, line spectral frequencies (LSFs) and multiple centroid analysis (MCA) coefficients as spectral representations. The metrics they used to calculate distance were: Euclidean, absolute, Kullback-Leibler and Mahalanobis distance, in addition to a method based on the Kalman filter. They tested three smoothing regimes: no smoothing, linear interpolation smoothing, and Kalman filter-based smoothing. They found the LSF spectral representation gave the best results, and that simple linear smoothing was preferable overall to both no smoothing and Kalman filter-based smoothing. Wouters and Macon [10] again used a similar approach of comparing a variety of the most common acoustic parameterisations (e.g. FFT- and LPC-derived cepstra and LSFs, both with and without various frequency-warping functions) and distance metrics (Euclidean distance, weighted Euclidean distance, Mahalanobis distance etc.). They concluded distance metrics using frequency warping performed better than those without, and that little advantage is conferred by using weighted distances or delta features. More generally, they concluded that the modest correlations observed with human ratings (0.66 maximum) indicates there is significant room left for improvement, and therefore the problem is far from solved.

Despite the fact that human speech is by its very nature articulated and generated within the vocal tract, only a small fraction of prior synthesis work (here we specifically mean concatenative and statistical parametric synthesis) has actually sought to exploit articulatory data as part of the synthesis process. As one relatively rare example, we have previously worked on including articulation into SPSS [14, 15], with the aim of both improving acoustic modelling and also introducing articulatory control over synthesis. In this paper, we propose articulation might also serve a useful purpose in unit selection synthesis, as an alternative to the usual acoustic join cost calculation. The motivating principle for this is simple: when a human speaks, their mouth is subject to physical constraints and cannot jump from one position to another. Therefore, when selecting units during Viterbi search for concatenation, we hypothesise it is reasonable to prefer a sequence of units which results in smooth evolution of mouth configurations. This paper aims to test this hypothesis. Specifically, using a large corpus of articulatory-acoustic speech data, we build three unit selection voices with the following join cost functions: i) standard acoustic baseline, ii) articulatory, and iii) combined articulatory and acoustic join costs. Note it is only the spectral subcost that is substituted, and the voices in all three cases are identical in every other way (including f0 and energy subcost functions, and their weighting in the total join cost). We use these voices to investigate how calculating join costs in the articulatory domain compares to a standard spectral join cost function, both in terms of the unit sequences selected and also in terms of subjective preferences in a listening test. As far as we know, this is the first work where an articulatory join cost is employed for unit selection.

We shall first describe our method in Section 2, and then present our results in Section 3. Finally, we shall summarise the conclusions we draw from these results in Section 4.

2. METHOD

This paper seeks to evaluate how an articulatory join cost compares to a standard spectral join cost by answering two core questions. First, we must establish whether and how the unit sequence that is selected when using an articulatory join cost differs from the standard voice. Second, assuming we indeed find significant differences, we then want to test whether there is any difference in listener preference between them.

We have used the `Festival Multisyn` [16] unit selection engine to perform these experiments, building voices using the `mngu0` [17] articulatory-acoustic data set. This data comprises over 1200 utterances with articulatory data recorded in parallel with the acoustic waveform using electromagnetic articulography (EMA) (a Carstens AG500 [18, 19]). The sentences contained in `mngu0` were selected from a large corpus of news text containing hundreds of thousands of sentences using a greedy text selection algorithm to ensure phonetic diversity. Therefore, though `mngu0` may be smaller than typical selection voice databases, it is the largest single-speaker articulatory corpus of which we are aware, and its speech data is in principle well-suited to unit selection voice building. We have used a subset of the articulatory data available in `mngu0`: x- and y-coordinates of 6 coils (upper and lower lips, jaw, and three points on the tongue) moving in the midsagittal plane.

The standard join cost in `Multisyn` has subcomponents for F0, energy and spectral distance. The spectral component uses Euclidean distance of 12 MFCCs. The MFCCs are z-score normalised during voice building, so this is equivalent to Mahalanobis distance.

In addition to the standard join cost, we implemented two further join costs which use the `mngu0` articulatory data: an *articulatory* join cost, and an *articulatory-acoustic* one. For the articulatory join cost, we substituted the spectral component for one which calculates Euclidean distance of the 12 EMA coordinates. We retained the F0 and energy subcomponents (and regardless refer to this as the "articulatory join cost" for convenience), since the EMA data contains no information pertaining to pitch or loudness. For the articulatory-acoustic one, we appended the 12 EMA coordinates to the MFCC vector used in the spectral subcost, again using Euclidean distance. In both cases, all features were z-score normalised during voice building. Apart from differences in the join cost calculation, the voices were identical in every way.

2.1. Unit sequence comparison

To compare the unit sequences that result from using the three join costs described above, we synthesised the 460 sentences contained in MOCHA TIMIT [20]. We chose this set of sentences since they were purpose-designed to be phonetically diverse, so ensuring the unit selection process would be applied to a wide variety of contexts. These sentences are also of a reasonably consistent and short length, making them suitable for use in the listening test below.

We compared the unit sequences synthesised using the three voices using two basic metrics. First, for a given pair of sentences, we counted the number of selected unit differences (**unit difference count**). This indicates the extent to which the different join costs affect which units are selected for concatenation. The second measure we used is the *join ratio*. In unit selection, units may be selected which were contiguous in the original utterance in the voice database, which means there is no real join between them. The **join ratio** is the number of true joins divided by the total number of possible joins (i.e. $N-1$, where N is the number of units) in the utterance.

2.2. Listener preference

We performed three preference tests to compare the three join cost types:

1. acoustic versus articulatory
2. acoustic versus articulatory-acoustic
3. articulatory versus articulatory-acoustic

We recruited 30 listeners, who were each presented with 25 pairs of stimuli for each of the three comparisons (though 1 participant failed to complete tasks 2 and 3 alas). The listening test was conducted in purpose-built perceptual testing booths using headphones and under controlled conditions. Participants were native speakers of English, mostly drawn from the student population, and were paid for taking part.

The stimuli for the test were drawn from the 460 TIMIT sentences synthesised as described in Section 2.1. We wanted to ensure only utterances with a good number of unit sequence differences were included in the preference tests. To achieve this, we looked at the per-sentence counts of unit differences between i) the acoustic join cost and the articulatory join cost, and ii) the acoustic join cost and the articulatory-acoustic join cost. We identified the sets of sentences in these two groups with greater than 70% units different from the sentences synthesised using the acoustic join cost. These sets contained 134 and 52 sentences respectively. We then found the intersection of these two sets (since we wanted to use the same sentences in each preference test), which contained 42 sentences. We ruled out some of these by inspection. First, we discarded sentences with errors in front-end processing, including: pauses inserted inappropriately by the phrasing module; faulty lexical entries or letter-to-sound rule application (e.g. words "corsage" and "mirage"); and inappropriate vowel reduction (e.g. weak verses strong forms of words such as "for" or "to"). In addition, longer sentences were less preferred. Using these criteria, we identified 25 utterances to test[1].

3. RESULTS

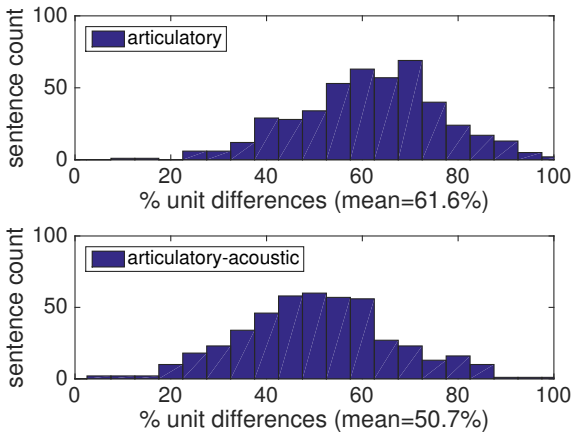

Fig. 1: Unit sequence differences as a percentage of the total number of units in the utterance. Differences are calculated with respect to the unit sequence obtained using the standard acoustic join for the articulatory join cost (top) and articulatory-acoustic join cost (bottom).

Fig. 1 summarises the unit sequence differences when using the articulatory and articulatory-acoustic join costs, compared to the acoustic join cost. This is presented as a histogram of per-sentence unit difference counts, expressed as a percentage of total number of units in each sentence. We see that on average 61.6% of selected units differ in the case of the articulatory join cost, and in some cases up to 100% of selected units are different. For the articulatory-acoustic join cost, this is somewhat less at an average of 50.7%, which is perhaps understandable. In both cases, the results confirm that the selected unit sequences differ significantly when using the articulatory join costs.

Fig. 2 explores the effect of articulatory join cost calculation on join ratio. For the articulatory join cost (top), we see that the join ratio tends to be increased. This means that a greater number of true joins are being made, which is a very interesting observation. This is not the case for the articulatory-acoustic join cost (bottom), which tends to give around the same, or perhaps slightly fewer, true joins.

Fig. 3 summarises the listening test results. We see both join costs which use articulation are preferred, with statistical significance, to the standard acoustic join cost (e.g. 58% in the case of the purely articulatory join cost). There is a small, though statistically significant, preference for the articulatory-acoustic join cost compared to the articulatory join cost, but this effect is weak.

[1]All stimuli and preference test responses available for download at `http://dx.doi.org/10.7488/ds/1315`

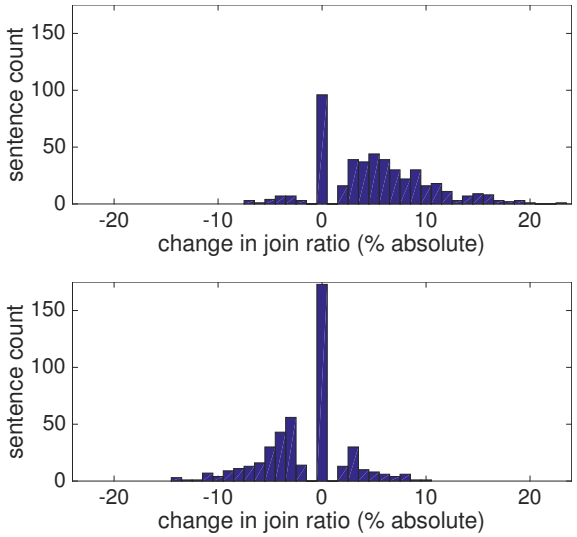

Fig. 2: Histograms of per-sentence change in *join ratio* (%-absolute change compared with the standard `Multisyn` acoustic join cost) for the **articulatory** join cost (top) and **articulatory-acoustic** join cost (bottom).

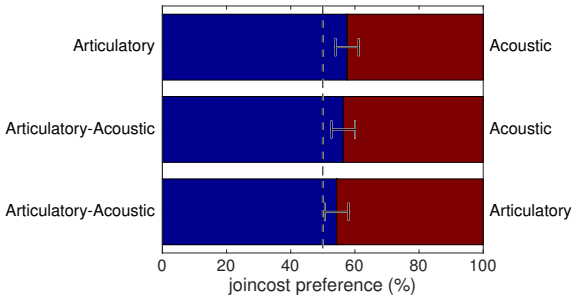

Fig. 3: Overall join cost preferences (with 95% CI)

Figs. 4 and 5 investigate preferences in more depth, showing per-sentence and per-listener preferences respectively. Fig. 4 shows, as is typical for unit selection, that the effect is not uniform, and while some sentences (the majority) are improved by the articulatory join cost, some sentences are perceived as worse. Fig. 5 shows listeners fall into roughly three groups. In the middle, there is a group who either cannot perceive any difference or have no clear preference for the articulatory join cost over the standard acoustic one. To the right, we see a group of listeners who can consistently distinguish them and tend to prefer the articulatory join cost. On the left however, we see one listener who seems to consistently prefer the standard acoustic join cost.

We have observed the articulatory join cost tends to give more frequent joins (Fig. 2), and is also significantly preferred to the acoustic one (Fig. 3). To better understand the relationship between these observations, we can look at how average listener preference varies with join ratio. Doing so for the preference test between the articulatory and baseline acoustic join costs, we find correlations of 0.45 and -0.35 respectively. So, while increasing the number of joins tends to result in worse listener scores for the acoustic cost, the opposite is true for the articulatory cost. This is a further very interesting difference between the two, with the behaviour of the articulatory cost arguably far more preferable. It must be noted though that these correlations are calculated on the basis of just 25 stimuli. In future, it will be imperative to test this observation again in a larger listening test.

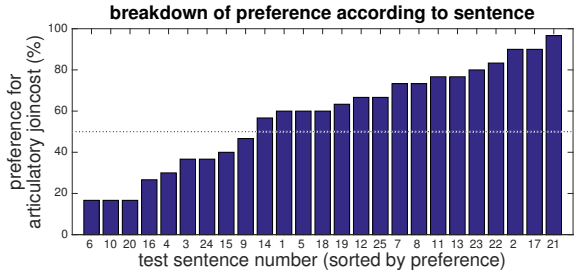

Fig. 4: Sorted per-sentence preference rates (articulatory cost)

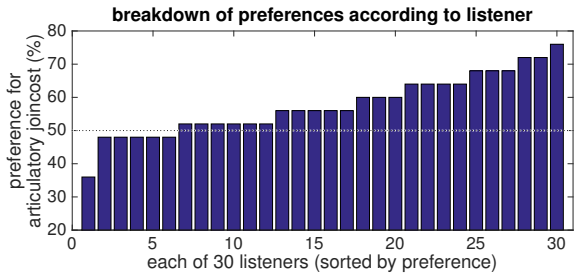

Fig. 5: Sorted per-listener preference rates (articulatory cost)

4. CONCLUSION

In this paper, we have proposed that computing join costs in the articulatory domain may be beneficial for unit selection synthesis. We have tested this by building voices using an articulatory-acoustic data set, and comparing three join costs: a standard acoustic join cost, an articulatory equivalent, and a combined articulatory-acoustic cost. By synthesising and analysing 460 TIMIT sentences, our results have confirmed that articulatory join cost calculation gives markedly different selected unit sequences. Preference test results have shown that the articulatory join cost was in fact preferred (58%), which is all the more interesting since our results also showed the articulatory join cost typically results in a significantly higher join rate in the synthetic speech. We conclude that articulatory join costs are promising, warranting further work in future. There is much that remains to be done, such as testing different join cost lengths, deltas, or weightings, and certainly to test other larger unit selection voices. For the latter, given the difficulty in recording large articulatory data sets, evaluating a voice for which articulator positions are not recorded but instead estimated from the acoustic signal alone ([21, 22]) is the highest priority in our view.

5. REFERENCES

[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[2] Alan W Black, "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling.," in *Proc. Interspeech*, 2006.

[3] Heiga Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7962–7966.

[4] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural network employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[5] Qiong Hu, Zhizheng Wu, Korin Richmond, Junichi Yamagishi, Yannis Stylianou, and Ranniery Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *Proc. Interspeech*, 2015.

[6] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank Soong, "TTS synthesis with bidirectional LSTM-based recurrent neural networks," in *Proc. Interspeech*, Singapore, September 2014, pp. 1964–1968.

[7] Alan W Black and Prasanna Kumar Muthukumar, "Random forests for statistical speech synthesis," in *Proc. Interspeech*, 2015, pp. 1211–1215.

[8] Vassilis Tsiaras, Ranniery Maia, Vassilis Diakoloukas, Yannis Stylianou, and Vassilios Digalakis, "Linear dynamical models in speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 300–304.

[9] Andrew Hunt and Alan Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, May 1996, vol. 1, pp. 373–376.

[10] Johan Wouters and Michael W Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis.," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1998, vol. 98, pp. 2747–2750.

[11] Esther Klabbers and Raymond Veldhuis, "Reducing audible spectral discontinuities," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 39–51, 2001.

[12] Yannis Stylianou and Ann K Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2001, vol. 2, pp. 837–840.

[13] Jithendra Vepa and Simon King, "Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1763–1771, 2006.

[14] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[15] Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, Jan. 2013.

[16] Robert A. J. Clark, Korin Richmond, and Simon King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[17] Korin Richmond, Phil Hoole, and Simon King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[18] Massimo Stella, Paolo Bernardini, Francesco Sigona, Antonio Stella, Mirko Grimaldi, and Barbara Gili Fivela, "Numerical instabilities and three-dimensional electromagnetic articulography," *Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3941–3949, 2012.

[19] Yana Yunusova, Jordan Green, and Antje Mefferd, "Accuracy assessment for AG500, electromagnetic articulograph," *Journal of Speech, Language and Hearing Research*, 2008.

[20] Alan Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.

[21] Korin Richmond, *Estimating Articulatory Parameters from the Acoustic Speech Signal*, Ph.D. thesis, The Centre for Speech Technology Research, Edinburgh University, 2002.

[22] Korin Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2835–2838.