ESTIMATING VELUM HEIGHT FROM ACOUSTICS DURING CONTINUOUS SPEECH

Korin Richmond
Centre for Speech Technology Research,
University of Edinburgh, UK
korin@cstr.ed.ac.uk

ABSTRACT

This paper reports on present work, in which a recurrent neural network is trained to estimate 'velum height' during continuous speech. Parallel acoustic-articulatory data comprising more than 400 read TIMIT sentences is obtained using electromagnetic articulography (EMA). This data is processed and used as training data for a range of neural network sizes. The network demonstrating the highest accuracy is identified. This performance is then evaluated in detail by analysing the network's output for each phonetic segment contained in 50 hand-labelled utterances set aside for testing purposes.

1. INTRODUCTION

Acoustic-to-articulatory inversion has occupied researchers for many years, which is unsurprising, since a successful method would find many applications: helping individuals with speech and hearing disorders by providing visual feedback; very low bit-rate speech coding; the possibility of improved automatic speech recognition (e.g. [7]) to name but a few.

Much early investigation was based on analytical techniques, such as inverse filtering, e.g. [4]. Later, articulatory synthesis models gained popularity as an aid to studying the inversion problem. More recent technologies such as X-Ray microbeam (XRMB) cinematography and electromagnetic articulography (EMA) have made it possible to record actual articulator movements in parallel with speech acoustics in a minimally-invasive way. This 'real' human data is arguably preferable to using analytical or synthesis techniques, where additional complications may be imposed by intrinsic flaws in the models themselves.

However, it seems relatively little work has been reported so far on the use of real articulatory data in deriving an acoustic-to-articulatory mapping. There are a few notable exceptions; in [2], EMA data was used to build a code-book of articulatory-acoustic vector pairs for 90 vowel transitions. [3] used XRMB data to train a simple supervised neural network to estimate the trajectories of three articulators for six English stop consonants.

The work presented in this paper can be thought of as extending the approach of [3]. A major difference is in the training data used. We report on the use of a corpus comprising just over 400 TIMIT sentences read by a male speaker, recorded at the EMA facility located in Edinburgh. This corpus embodies much greater phonetic diversity than that found in the rather restricted data set at the disposal of [3]. Thus, in principle, there is more scope for investigating the one-to-many problem many researchers have cited as inherent to an inversion mapping. For example, while many researchers have demonstrated how different articulatory configurations can produce acoustic signals which are indistinguishable from each other (e.g. articulatory compensation in bite-block experiments), the exact extent to which this affects unconstrained continuous speech remains uncertain. Furthermore, it is sensible to question whether this factor might be language- or even speaker-dependent. Parallel acoustic-articulatory data, such as that provided by EMA, obviously offers considerable hope that such questions may ultimately be settled.

Another key contrast with [3] is that velum height was chosen as the articulatory parameter to be estimated from the acoustic signal. Nasals can exhibit notoriously more opaque spectral effects than segments produced without nasal coupling, which again represents more of a challenge. Furthermore, methods to estimate degree of nasalisation during speech have in themselves been pursued by several researchers, e.g. [1].

This paper will first outline the steps involved in processing the 'raw' data, with particular comment on difficulties regarding normalisation of EMA data which became evident. The task of training neural network models using this data is then briefly covered. Next, we describe the method employed to scrutinise more closely the best performing networks. Examples of the findings of this method are included for discussion, focusing equally on areas where the network performs reasonably well and areas where the network performs poorly.

2. DATA PROCESSING

The steps taken to prepare the raw acoustic and articulatory data for training a neural network are very similar to those described in [3]. Briefly, the process may be outlined as follows:

1. A simple endpoint detection algorithm, using log power and a user defined threshold, isolates the start and finish of the utterance within the recording. This is to avoid varied articulator movements during silent stretches adversely affecting supervised learning.

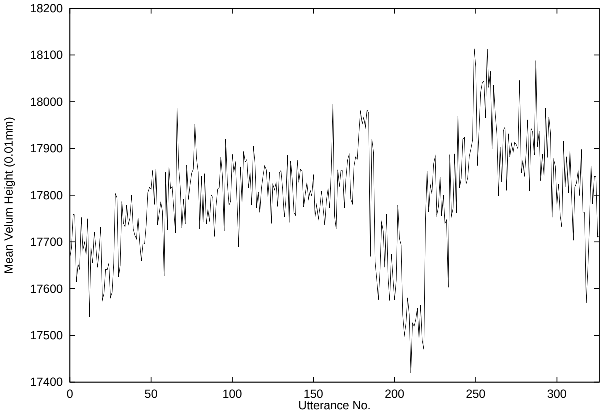

Figure 1: Example of a corpus demonstrating poor consistency of mean velum heights throughout.

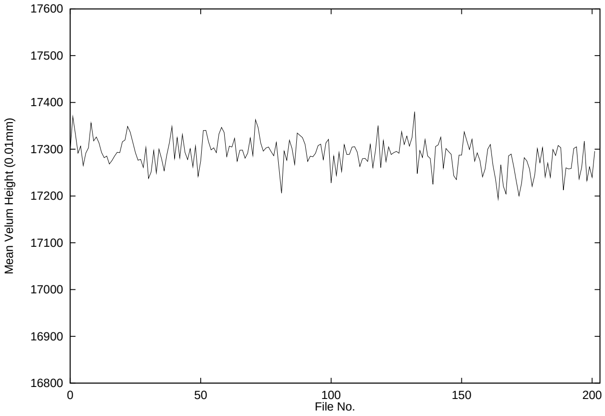

Figure 2: Plot of means of the corpus used in the present paper (each file contains two utterances).

2. The waveform, sampled at 16 kHz with 16 bit precision, is segmented into overlapping frames of 16ms duration with 8ms shift.

3. Filterbank analysis is carried out, yielding 16 filterbank coefficients for each time frame.

4. The corresponding subsection of the EMA velum height trace is resampled to match the frameshift of the acoustic coefficients (8ms).

5. The data is normalised and converted to the correct format for network training. This is described in greater detail below.

2.1. Data Normalisation

The acoustic and articulatory coefficients need to be normalised to ranges more suitable for neural networks. Specifically, the filterbank coefficients were normalised to fall between 0.0-1.0, while the velum height target values were normalised to between 0.1 and 0.9. The EMA velum height and 16 filterbank coefficients at each time point were normalised using their respective means and standard deviations. In order to ensure consistency between the training and testing data sets, the necessary means and standard deviations were calculated over the whole corpus.

It seems prudent to mention a problem apparently inherent to EMA data which can make it difficult to normalise consistently across a whole corpus of utterances. Figure 1 shows a plot of the mean velum heights for each utterance from an earlier recording session. The mean velum height should vary moderately, corresponding to the phonetic content of the utterance. For example, if an utterance contained a significant number of nasal segments, we would expect the overall mean velum height for that utterance to be lower than an utterance which contained a significantly high number of oral stops and no nasal segments. Hence, we assume we should see a reasonably consistent range, or band, of mean velum heights across a given corpus, within which the values for specific utterances will vary randomly. However, the range of variation seen in Figure 1 is excessive, and no doubt reflects facets of EMA-recorded data.

Perhaps the two most detrimental aspects evident in Figure 1 are *discontinuities* and *drift* in the means. Discontinuities (a sudden shift in the apparent location of the band of expected variation) can result from events during recording such as when a receiver coil becomes detached from its articulator and is then replaced in a slightly different location. Drift (a gradual elevation or declination of the band of expected variation) is hypothetically attributable to various sources. It may be the case that the speaker grows more and more accustomed to the presence of the EMA coils on the articulators over time, and modifies articulation accordingly. If true, this could prove difficult to avoid, as it seems impractical for the speaker to wear the EMA coils for any great length of time prior to recording. Another possibility is that a temperature drift in the EMA system may in turn result in a drift in the absolute values recorded [5].

Obviously, inconsistencies of any type will pose a problem for neural net learning, but unfortunately it is not immediately clear what a satisfactory remedy might be. Moreover, the problems mentioned so far relate to a single speaker within a single recording session; additional challenges can be expected to arise when normalisation across different speakers becomes necessary. For example, different speakers will have differently shaped heads and articulatory tracts, with potentially different ranges of articulatory movement.

In the meantime, however, the characteristics of the data used for the purpose of this paper are shown in Figure 2. As can be seen, the mean velum heights are reasonably consistent with no serious discontinuities evident (after removing a few outliers by hand). There is perhaps a very slight downward drift towards the end of this data set. However, this did not result in the problems encountered in previous studies, where discontinuities and severe drift made it impossible for the network to learn appropriate output ranges.

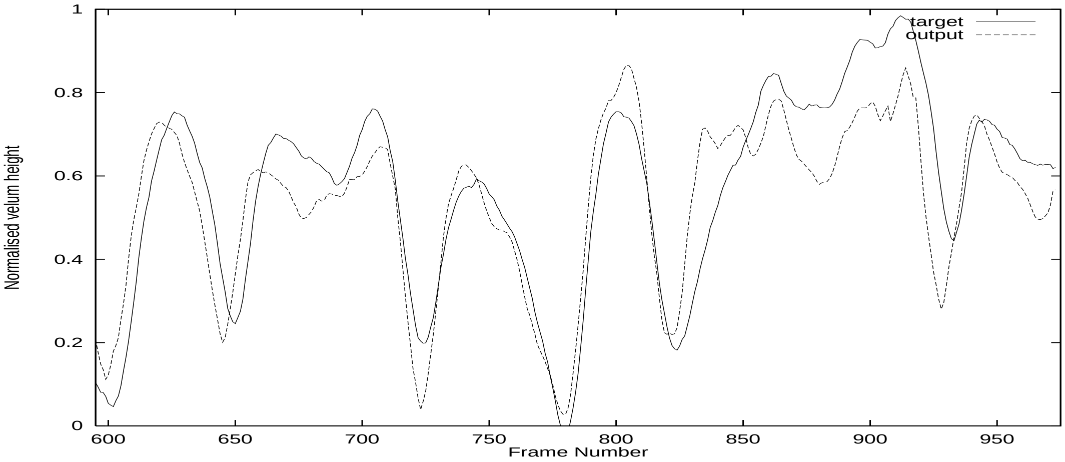

Figure 3: An example of network estimated velum height compared with the actual velum height trajectory. The troughs at approximate frame numbers 600, 650, 725, 775, 825, and 925 correspond to the nasals in the TIMIT sentence "pla**NN**ed pare**N**thood orga**N**isatio**N**s pro**M**ote birth co**N**trol" respectively.

3. NETWORK TRAINING

The neural networks were modelled using the Stuttgart Neural Network Simulator (SNNS) package, V4.1 [1]. Therefore, the data had to be processed according to the steps outlined in Section 2 and converted into the appropriate SNNS pattern file format. This was performed by purpose-coded utilities written in C++, taking advantage of CSTR's "Edinburgh Speech Tools" class library.

The total corpus of 404 utterances was divided into two groups: one for testing (100 utterances) and the other for training. The files set aside for testing purposes were selected at random, so that the removed files spanned the whole of the corpus. This was to account for any inconsistencies still remaining within the corpus.

Similar to [3], a supervised neural network with a large "context" input window of 25 time frames (400 input units, as there are 16 filterbank coefficients for each frame), two hidden layers, and a single output unit was used. A key difference was the introduction of recurrence by adding Elman-style context units, with disabled self-recurrent links, for the second hidden layer. We found this not only resulted in decreased overall training time, but also in much smoother output trajectories from the trained network.

Networks with various numbers of units in the two hidden layers were tried and evaluated using the test set (to avoid "fitting" to the training set) as the basis of a simplistic, if computationally expensive, search strategy for a reasonable network. With less than 16 units in the first hidden layer, it was found the network would be unstable when training and would not converge satisfactorily. However, adding just one extra unit in the first hidden layer adds at least an extra 400 links with the input layer that must be updated. This obviously has a dramatic impact on training speed. Realistically, there exists a certain trade-off between network accuracy and reasonable training time. Networks were tried with the first hidden layer varying between 16 and 20 units and the second layer ranging from 12 to 16 units. The network demonstrating the best performance so far has 20 units in the first hidden layer and 14 in the second.

4. NETWORK EVALUATION

Once the best trained networks had been identified, they could be fed novel test utterances to give a qualitative impression of performance, as in Figure 3. This plot gives a relatively concise representation of typical performance. Both the target and the network-estimated trajectories have been smoothed using a basic five point average window.

As Figure 3 shows, even where the network has correctly detected the presence of a velum-lowering gesture, it may still be rather poor at estimating the exact timing or magnitude of that gesture. For example, the lowering gesture estimated roughly at frame number 725 is in the right place, but of excessive magnitude. Further investigation will focus on trying to ascertain why this might be the case.

Qualitative examination of estimated velum height trajectories suggests that the trained network is not able to estimate the velic gestures corresponding to different phonetic segments with equal accuracy. We maintain it will be fruitful to explore this impression more quantitatively.

Two complementary numerical measures were used in [3] and [6] to evaluate accuracy of network-estimated velum height compared to the actual velum height trajectory; root mean square-error (RMSE) and Pearson product-moment correlation (PPMC), taken from [6] as:

$$r = \frac{N\sum_{p=1}^{N} o_p d_p - \sum_{p=1}^{N} o_p \sum_{p=1}^{N} o_p}{\sqrt{\left[N\sum_{p=1}^{N} o_p^2 - \left(\sum_{p=1}^{N} o_p^2\right)^2\right]\left[N\sum_{p=1}^{N} d_p^2 - \left(\sum_{p=1}^{N} d_p^2\right)^2\right]}}$$

where o and d are vectors of actual and estimated velum height trajectories. RMSE indicates the distance of the

[1] http://www.informatik.uni-stuttgart.de/ipvr/bv/projekte/snns/snns.html

Seg.	# found	# frames	RMSE	RMSE-mean	RMSE-sd	PPMC-mean	PPMC-sd
n	79	822	0.2039	0.1684	0.0989	0.4421	0.2673
ng	10	97	0.1882	0.1704	0.0906	0.3616	0.2871
a	27	333	0.1491	0.1285	0.0822	0.3498	0.2886
ai	19	317	0.1488	0.1336	0.0645	0.3053	0.1883
ch	14	164	0.1430	0.1107	0.0734	0.2433	0.2148
m	37	351	0.1424	0.1166	0.0755	0.3971	0.4803
f	23	257	0.1248	0.1003	0.0727	0.2436	0.2939
s	82	1056	0.1179	0.1050	0.0686	0.2524	0.3015
jh	8	97	0.1153	0.0956	0.0612	0.1805	0.1421
v	12	109	0.1034	0.0850	0.0631	0.2310	0.2285
z	13	124	0.0886	0.0841	0.0384	0.2058	0.0899

Table 1: RMSE and PPMC figures for a selection of segments

output trajectory from the actual trajectory, while PPMC is a measure of correlation between output and target trajectories, i.e. similarity of trajectory ‘shape’. Obviously, calculating these measures over an entire utterance containing twenty or thirty segments will not provide much insight. It would be more helpful to be able to calculate RMSE and PPMC measures for each separate segment in an utterance. The values for one segment could then be combined with values for the same phone in other utterances.

To implement this idea, fifty test utterances were hand labelled. A utility program was written which would take each segment in these label files in turn and calculate its RMSE and PPMC. In this way, statistics pertaining to RMSE and PPMC could be compiled for each phone in the labelling phone list. Specifically, the overall RMSE, RMSE mean (i.e. the average of the RMSE’s for each instance of a phone) and standard deviation, as well as the mean and standard deviation for the PPMC of each phone, were calculated; see Table 1, which shows example statistics for selected segments.

DISCUSSION

As hypothesised, the network tested here did not estimate velum height for all segments with equal accuracy. Interestingly, the nasal stops scored the highest RMSE, while their PPMC were also among the highest. This would indicate the network is generating an appropriate gesture in roughly the right places, but that it does relatively poorly at estimating the magnitude of the gesture. Visual inspection of the output tends to supports this view, as discussed in Section 4. The velum height for the low back vowel (for which the velum is also presumed to lower), is similar in nature to that for the nasal stops. An additional point worth noting is that the voiceless affricate and fricatives all have higher RMSE than their voiced counterparts. Further investigation will be required in order to understand why specific phone-dependent differences are observed.

Finally, it is important to point out that networks with different architectures or from different training runs were typically found to produce a similar ranking of the phones in terms of RMSE and PPMC, which would suggest that such observations are grounded in more than just the random dynamics of neural network training.

CONCLUSIONS

This paper has reported on an experiment which endeavours to train a neural network to perform an inversion mapping for phonetically rich speech, and includes reference to EMA normalisation difficulties that became evident in the process. The method currently being used to analyse the performance of a such a network has been described. This phone-by-phone “breakdown” has already shown promise at providing useful insight. Further work will concentrate on using this method to investigate how the observed characteristics of the network performance at mapping from acoustics to the articulatory domain might be explained.

REFERENCES

[1] M. Chen. Acoustic parameters of nasalised vowels in hearing-impaired and normal-hearing speakers. *J. Acoust. Soc. Am.*, 98(5):2443–2453, November 1995.

[2] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman. Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *J. Acoust. Soc. Am.*, 100(3):1819–1834, September 1996.

[3] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy. Inferring articulation and recognising gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, 92(2):688–700, August 1992.

[4] H. Wakita. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Trans. Audio Electroacoust.*, 21:417–427, 1973.

[5] Alan Wrench. Personal communication. (Queen Margaret College University, Edinburgh).

[6] J. Zachs and T. R. Thomas. A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech and Language*, 8:189–209, 1994.

[7] I. Zlokarnik. A speech recognizer using electromagnetic articulography. (English short version of PhD thesis), 1995.