Continuous Speech Recognition Using Articulatory Data

Alan A. Wrench and Korin Richmond

Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh

CSTR, University of Edinburgh

ABSTRACT

In this paper we show that there is measurable information in the articulatory system which can help to disambiguate the acoustic signal. We measure directly the movement of the lips, tongue, jaw, velum and larynx and parameterise this articulatory feature space using principal components analysis. The parameterisation is developed and evaluated using a speaker dependent phone recognition task on a specially recorded TIMIT corpus of 460 sentences. The results show that there is useful supplementary information contained in the articulatory data which yields a small but significant improvement in phone recognition accuracy of 2%. However, preliminary attempts to estimate the articulatory data from the acoustic signal and use this to supplement the acoustic input have not yielded any significant improvement in phone accuracy.

1. INTRODUCTION

There have been a number of studies in the last 10 years which have investigated the potential of directly measured speech production parameters to improve the accuracy of automatic speech recognition systems (ASR) [1].

Zlokarnik [2] used an HMM-based speech recognition system that made use of simultaneously recorded acoustic and articulatory data, gathered by means of Electromagnetic Articulography (EMA). The data described the movement of small coils fixed to the speakers' tongue and jaw during the production of German V1CV2 sequences. The coordinates of the coil positions, their first derivatives, mel cepstra and acoustic energy were weighted according to their ability to discriminate between phonemes and concatenated in various combinations to form acoustic/articulatory feature vectors. These acoustic and articulatory feature vectors were evaluated for two subjects (one male and one female) on a speaker-independent isolated word recognition task. When the articulatory measurements were used as input on their own, the word error rate increased by a relative percentage of 300%. The recognition rate dropped from 85.8 using the acoustic input to 56.7% using the coil positions and their first derivatives. However, the discriminant power of the combined representation was capable of reducing the error rate of comparable acoustic-based HMMs by a relative percentage of more than 60%. The recognition rate rose from 85.8 using the acoustic input to 94.8%.

Soquet et al [3] used a larger corpus (1536 CVCVs vs. 165 VCV's), but did not measure tongue movement data, relying instead on 3 EPG contact coefficients (anterior, posterior and centrality). The Movetrack articulography system provided upper lip, lower lip and jaw movement data and this was supplemented with air pressure measured within the oral cavity. The articulatory data performed poorly on its own (36.8%), but when combined with the acoustic data the word recognition rate rose from 44.6 to 83%.

These results provide a basis for optimism, however, the recognition tasks are simple and the baseline system performances are not state-of-the-art. It is well understood that the better the baseline recogniser performance, the harder it is to make gains. Speaker independent recognition and continuous speech recognition using directly measured data remain to be seriously tested. In the project introduced in this paper, we hope to extend the promising work of Zlokarnik. Firstly, by creating a database with the additional articulatory information provided by an Electropalatograph (EPG), Laryngograph and EMA. Secondly, by using a corpus which represents English read speech, incorporating a broad coverage of co-articulation in sentence structures. Thirdly, by using a baseline speaker independent continuous speech recognition system tuned to provide state-of-the-art performance before comparisons between acoustic and articulatory feature vectors are made.

2. DATA AND BASELINE

2.1. Database

The MOCHA (Multi-CHannel Articulatory) database used in this paper is being created to provide a resource for training speaker-independent continuous ASR systems and for general co-articulatory studies. The planned dataset includes 40 speakers of English, each reading up to 460 TIMIT sentences (British version). The articulatory channels currently include Electromagnetic Articulograph (EMA) sensors directly attached to the vermilion border of the upper and lower lips, lower incisor (jaw), tongue tip (5-10mm from the tip), tongue blade (approximately 2-3cm posterior to the tongue tip sensor), tongue dorsum (approximately 2-3cm posterior to the tongue blade sensor) and soft palate (approximately 10-20mm from the edge of the hard palate). A Laryngograph provides voicing information and an Electropalatograph (EPG) provides tongue-palate contact data at 62 points across the hard palate. Acoustic data was recorded simultaneously with these articulatory measures.

For this paper two speakers - one male and one female - were analysed. Phonemic transcriptions were generated automatically for each speaker using a single entry keyword dictionary [1]. Rules, applied to the keyword dictionary generate a dialect-dependent dictionary. The orthography is then transcribed using this dictionary with post-lexical rules to take care of word boundary effects such as r-sandhi. Transcription errors are estimated to lie between 5 and 10 percent causing a corresponding reduction in ASR phone accuracy measures across the board in this paper. Work is underway to refine these transcriptions.

2.2. Baseline system

The ASR training and testing was performed using a jackknife procedure where test group 1 consists of 92 sentences numbered 1,6,11… from the corpus; test group 2 consists of 2,7,12…; etc. with the remaining 4/5ths of the sentences in each case used for training. Training and testing was carried out in this way so that recognition accuracy (NIST) scores were generated for all 460 sentences.

The baseline system was generated using HTK v2.1[1]. Acoustic features were mel-scaled cepstral coefficients using 24 filterbank channels based on a 16KHz sampled 16bit speech signal with a hamming windowed frame of 25ms sampled every 10ms. The number of cepstral coefficients was varied from 12 to 18 in steps of 2 and cepstral liftering was used. A normalised energy measure was added; computed as the log of the signal energy divided by the maximum frame value for the utterance. Delta and deltadelta coefficients were calculated using 2[nd] order recursion and appended to the feature vector producing vectors of lengths 39,45,51 and 57.

The HMMs were implemented as left-to-right models with 3 states. Output probabilities were modelled by between 2 -7 mixtures of Gaussian probability density functions (PDFs). A phone bigram was trained using all 460 sentences. 57 monophones models were trained from a flat start and cloned to produce approximately 5500 triphones. Following re-estimation, a decision tree was used to tie states and 101,614 logical models were synthesised using between 5700 and 7000 physical models. Insertion penalty (1.0) and bigram weight (8.0) were optimised to maximise accuracy scores on the first jackknife test set and found to be the same for both speakers.

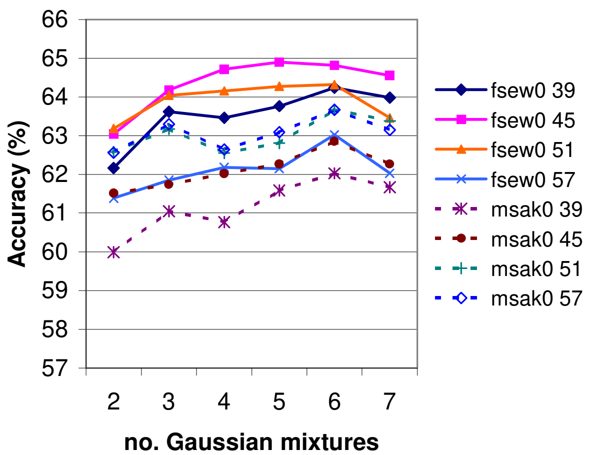

Figure 1: Accuracy for baseline acoustic system using N mfcc coefficients + 1 energy coefficient + deltas + deltadeltas input features. Trials for N=12,14,16 and 18 mfccs are shown for the female speaker fsew0 and male speaker msak0.

Figure 1 shows the mean accuracy scores across all 460 sentences for 12,14,16,18 cepstral coefficients and for 2-7 Gaussian mixtures.

2.3. Baseline results

It should be noted that the best performance for fsew0 was with 14 mfccs and for msak0 with 16 or 18 mfccs. The higher number of coefficients for the male speaker is probably due to the formants being closer together for males.

3. ARTICULATORY FEATURES

In order to use the raw articulatory data as input to an ASR system the disparate sources must be combined, correlated components should be removed, and the dimensionality must be reduced. There are many ways to achieve this. As a first attempt, we have used principal components analysis (PCA).

3.1. EMA data

EMA data consists of x and y co-ordinates for upper and lower lip, jaw, 3 tongue locations and velum making 14 coefficients in total, sampled at 500Hz. First of all this data was downsampled to 100Hz, channel by channel. Then, the velocities and accelerations associated with these displacements were added to make a 42 dimensional vector. PCA was applied to reduce this vector to either 24, 30, 36 or 42 dimensions.

3.2. EPG data

EPG measures tongue/palates contact over the whole palate. Data consists of 62 on/off values per frame sampled at 200Hz. PCA was applied to every second frame to reduce this to a 4 dimensional feature vector sampled at 100Hz. Some more detail on this process can be found in Wrench &Hardcastle [4].

3.3. Laryngograph data

The Laryngograph data measures the change in glottal contact at high frequencies thus providing pitch and voicing information. The signal is recorded at 16kHz. To produce a measure of voicing energy at 100Hz, the signal was differentiated and the root mean square of non-overlapping, 160-sample frames was calculated.

3.4. Articulatory feature vector

The combined articulatory feature vector was built by taking the first 4 EPG principal components along with the voice energy value and adding the corresponding deltas and deltadeltas to create a 15 dimensional vector. This was concatenated with 30, 36 or 42 principal components derived from the EMA data and to result in an overall vector size of either 45, 51 or 57. PCA was applied again in each case to diagonalise the covariance matrix without further reducing the dimensionality.

3.5. Results

The best performance for speaker fsew0 is 55% accuracy, achieved using 30 EMA principal components (figure2).

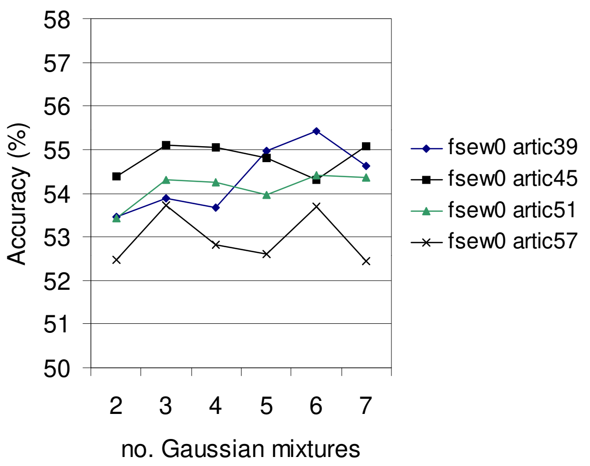

Figure 2: Mean accuracy results for fsew0 with articulatory feature vector sizes of 39, 45, 51 and 57 corresponding to an increase in the number of EMA components from 24 to 42.

4. ARTICULATORY AND ACOUSTIC COMBINED

The combination of acoustic and articulatory features was again achieved by applying principal components analysis to reduce the dimensionality and diagonalise the covariance matrix. The best acoustic vector was concatenated with size 39 and 45 articulatory vectors and PCA was then applied resulting in vectors of size 45, 51 and 57.

Figure 3 compares the best performing acoustic baseline with the different combinations of articulatory and acoustic vectors. The measurement confidence limits indicate a significant improvement in accuracy. Figure 4 shows a similar improvement for speaker msak0.

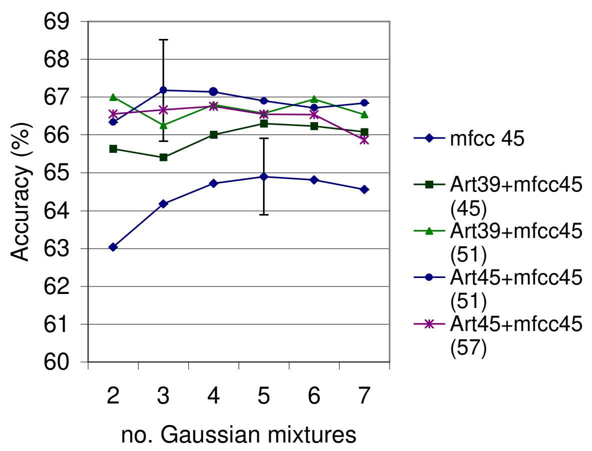

Figure 3: Mean accuracy results for fsew0. Combination of articulatory feature vector with acoustic vector using principal components analysis to reduce the dimensionality. E.g. art45+mfcc45 (51) starts with a 90D feature vector and results in a 51 dimensional vector. 95% confidence limits are marked.

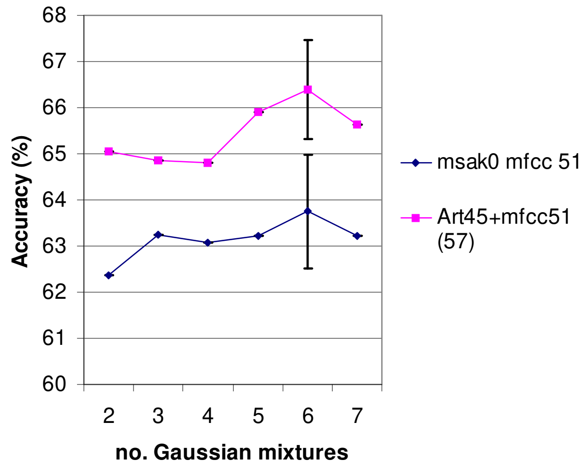

Figure 4: Mean accuracy results for msak0. Combination of articulatory feature vector with acoustic vector using principal components analysis to reduce the dimensionality. E.g. art45+mfcc51 (57) results in a 51 dimensional vector. The 95% confidence limits are shown for 6 mixture case.

5. ESTIMATING ARTCULATORY DATA FROM ACOUSTIC DATA

Having shown that articulatory features can enhance recognition accuracy, the challenge is to transfer this advantage to the practical circumstances where only acoustic data is available. To do this we have employed a multilayer perceptron with 2 hidden layers of size 50 and 25. The input to this network consists of 20 time-shifted (10ms) sets of 24 mel-scaled filterbank coefficients sampled from a 25ms hamming windowed. The output is the 14 EMA channels and the first 8 EPG principal components. Using these estimated values has not so far resulted in any improvement in accuracy over the baseline acoustic results. More details showing typical root mean square error and mean correlation values for a similar estimation procedure can be found in Frankel et al [5].

6. DISCUSSION

The speaker dependent continuous speech tests shown here show that there is a gain in segment level recognition accuracy although it is nowhere near as marked as the performance improvements determined at word level by Zlokarnik [1] and Soquet [2]. We have optimised the acoustic recognition parameterisation and then used the same baseline system simply swapping the input feature vector. This should, if anything, bias the results in favour of the acoustic input. However, the improvement in accuracy is shown to be significant and of the same order as can be achieved by increasing the number of mixtures from 2 to 6 in the baseline system. Although we have not presented the results here we also checked whether the application of PCA to the acoustic input on its own could account for the improvement but this was found to make no significant difference.

The business of processing articulatory data is more complicated than acoustic processing and it seems likely that the small improvement demonstrated in this paper can be expanded upon.

6.1. Measuring additional articulatory parameters

There are advances still to be made in the measurement of articulatory data. Notably, the measurement of the glottal opening gesture could provide a key source of data for distinguishing voiced from voiceless consonants - a significant source of segmental confusion in the baseline system. (NB. The glottal opening gesture is a low frequency signal not picked up by the Laryngograph)

6.2. Trying different combinations of measurements.

We have carried out only one possible arrangement for combining the articulatory data streams. It is unlikely to be the optimum approach.

Principal components analysis as a tool for combining the feature vectors, reducing the dimensionality and diagonalising the covariance matrix generated by the resulting vector is convenient but fairly crude. It does not take into account the relative value that the input data channels have for discriminating between phone classes. This can be addressed by using linear discriminant analysis applied to carefully selected classes.

6.3. Transformation of measurements prior to combination

We have used raw EMA positional data in this study. However, the mapping from articulatory configuration to the acoustic output depends upon the shape of the vocal tract cavities and not on absolute positions. It is well known, for example, that substantially different jaw and tongue combinations can produce very similar vocal tract cavity structures. Nonlinear transformation of the geometrical measurement space can help to reduce this confusion. This may improve performance, particularly where there is limited training data. For example referencing tongue sensors to the hard palate would be one transformation which might help in the future not only to reduce the many-to-one mapping redundancy but also to normalise data from different speakers.

6.4. Alternative modelling approaches

Work done in the field of audiovisual speech recognition indicates [6] that when using the HMM engine it may be more beneficial to combine the articulatory and acoustic streams later in the recognition process. However, because the articulatory parameters change smoothly over time, linear dynamical modelling may be a more appropriate paradigm for this kind of data [5][7].

7. CONCLUSION

We have shown that the accuracy of a state-of-the-art speaker-dependent continuous-speech ASR system can be enhanced by adding directly measured articulatory data. The percentage improvement attained in this paper is about 2%. This is equivalent to an error reduction of about 6%. The methods used to achieve this improvement have been fairly basic and it is therefore very likely that further substantial gains in performance can be achieved. Preliminary trials with articulatory data estimated from the acoustic signal have not as yet resulted in an improvement in phone recognition accuracy. We will continue to work on improving the articulatory parameterisation. and the estimation procedure.

8. ACKNOWLEDGEMENTS

This study was supported by EPSRC grant GR/L78680.

9. REFERENCES

1. Wrench, A.A., "A multi-channel/multi-speaker articulatory database for continuous speech recognition research", *In Phonus, Research Report No. 4*, Institute of Phonetics, University of Saarland, In press, 2000.

2. Zlokarnik, I., "Adding articulatory features to acoustic features for automatic speech recognition*", Acoust. Soc. Am. 129th Meeting*, Abstract 1aSC38, 1995.

3. Soquet, A., Saerens, M, and Lecuit, V, "Complimentary cues for speech recognition", *Proc. Int. Conf. Phonetic. Sci.*, 1645-1648, 1999.

4. Wrench, A.A. and Hardcastle W. J., "A multichannel articulatory speech database and its application for automatic speech recognition.*", Proc. 5th seminar on speech production: models and data*, 305-308, 2000.

5. Frankel, J., Richmond, K., King, S. & Taylor, P., "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces.", *6th Int. Conf. Spoken Lang. Proc.*, In press, 2000.

6. Dupont, S. and Luettin, J., "Using the Multi-Stream Approach for Continuous Audio-Visual Speech Recognition: Experiments on the M2VTS Database*", 5th Int. Conf. Spoken Lang. Proc.* , CDROM, 1998.

7. King, S. and Wrench A., "Dynamical system modelling of articulator movement.", *Proc. Int. Conf. Phonetic Sciences*, 3, 2259-2262, 1999.