A NEW RESOURCE FOR PRODUCTION MODELLING IN SPEECH TECHNOLOGY

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

1 INTRODUCTION

1.1 Overview

In the field of speech technology research, observation and analysis of speech has been limited to one signal: The acoustic waveform. In a noise-free environment the acoustic waveform is all that a mature human listener requires to comprehend speech. It is also the most practical signal to record and so will remain the primary input and output mode for speech technology. However, in this paper we argue that an articulatory representation of speech provides a comprehensive description of speech which is not only equal to that of the acoustic signal but also contains additional information which is not present in a standard mel cepstrum based acoustic representation.

Datasets containing directly measured articulatory time series suitable for training automatic speech recognition (ASR) systems are difficult to create and have until recently been unavailable. The MOCHA (MultiChannel Articulatory) database is designed to address this need. This paper goes on to discuss in detail the potential role of articulatory time series in the training of an ASR system. Concrete examples are presented showing the performance equivalence of articulatory and acoustic signals using a state-of-the-art HTK-based ASR system. These examples show improvements on previously published results [Wrench & Richmond, 2000] providing supporting evidence for the value of this resource to the speech technology community.

1.2 The Rationale for an Acoustic/Articulatory Dataset

There has been growing interest within the ASR community in using articulatory parameters, either as a supplement to or substitute for spectrally based input parameters. There has been a longstanding request [Rose et al, 1994] by the ASR community for a combined articulatory/acoustic corpus suitable for the task of speaker-independent continuous speech recognition.

1.2.1 A Linguistic and Biomechanical Viewpoint

One key problem in the design of speech recognition systems is coping with the high level of inter- and intra-speaker variability. Variability occurs at the phonological and phonetic levels. At the phonetic level this is primarily due to the physical structure of a speakers passive (e.g. hard palate) and active (e.g. tongue) articulators and to the dynamics of the active articulators given the physical constraints and neuromuscular settings adopted by the speaker to generate sound pattern sequences. Some theorists [Browman & Goldstien,1992] in fact go further and link articulatory gestures to the phonological level. Studies supporting this Articulatory Phonology theory [Jung et al 1996] suggest that variation in the extent and timing of articulatory gestures can account for many segmental deletions and assimilations commonly encountered in casual speech. This provides a theoretical basis for supposing that articulatory parameters could prove to be more robust to inter- and intra-speaker variability.
1.2.2 Experimental Evidence

An articulatory feature space is constrained by biomechanical properties and therefore we would argue that has the potential to improve the accuracy when training an automatic speech recognition (ASR) system; form a basis for picking units for synthesis; or expose idiosyncratic sequences for speaker recognition.

Previous research in applying speech production modelling to automatic speech recognition has largely involved inferring articulatory features from the acoustic data using vocal-tract models [Schmidbauer et al, 1993][Ramsay, 1998][Richards et al, 1995] or linguistic rules [Deng & Erler, 1992][Deng & Sun, 1994][Kirchoff, 1996][Kirchoff, 2000]. Results have been promising but inferring articulatory parameters by such means is subject to errors in labelling of segments; in the theory of phonetic features which make up these segments and in the assumption that these features are binary and synchronise with the phonetic boundaries.

There have been some experiments using restricted articulatory datasets [Papcun et al, 1992][Zlokarnik et al, 1995][Jung et al, 1996][Zacks, 1994][Zlokarnik, 1995][Soquet et al, 1999][King & Wrench, 1999] which show improvements for limited recognition tasks such as nonsense words.

1.2.3 Articulatory Datasets

Directly measured multi-channel articulatory datasets in large enough quantities to perform statistical analysis from which speaker-independent conclusions can be drawn are rare due to the difficulty of keeping sensors attached to the tongue and soft palate and the cost of purchasing and running the measurement instrumentation. The only other large publicly available kinematic speech corpus was created at the Wisconsin x-ray microbeam facility. The collection consists of 60+ American English speaker datasets. Each dataset contains a set of tasks including: two prose passages (13%); counting and digit sequences (6%); oral motor tasks (8%); citation words, near-words, sounds and sound sequences (33%) and sentences (40%). The sentences consist of 21 TIMIT sentences and 19 other sentences with varying numbers of repetitions. The MOCHA-TIMIT dataset differs in 5 key aspects.

1. The corpus consists of 460 phonetically-compact sentences designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Some extra sentences have been added to the 450 original American English TIMIT (sx) sentences. These extra sentences capture phonetic pairs and contexts which occur in the RP accent of British English.

2. The dataset contains speakers who speak many dialects of English including British, American and non-native. Some speakers may have phonetic pairs which are not represented by the sentence set (e.g. Scottish dialect) but the majority of phonetic contexts are represented.

3. The recordings are made in a sound-damped studio providing higher quality audio than is possible with the microbeam apparatus used for the Wisconsin dataset.

4. The MOCHA dataset has extra articulatory channels corresponding to the soft palate, tongue-palate contact (EPG) and vocal fold movement (EGG) which supplement the lip, tongue and jaw channels present in the Wisconsin dataset.

5. The MOCHA database is labelled at the phone level by an automatic forced alignment procedure.
2. A MULTICHANNEL ARTICULATORY DATASET

2.1 Edinburgh Speech Production Recording Facility

In 1995 a speech production facility was set up in Edinburgh with one of the main goals being the recording of a large multi-channel multi-speaker articulatory database. The facility consists of a purpose-built sound-damped studio and control room with a Carstens AG100 Electromagnetic Articulograph system, a laryngograph (later superseded by a Rothenberg Electroglottograph (EGG) system), an Electropalatograph (EPG) and a microphone. The EGG and microphone signals are recorded as two channels directly onto computer through a digital sound card installed in a PC. Another two PCs are used to record the EMA and EPG data directly. The three systems are synchronised using serial port communication combined with signal post-processing.

2.2 The MOCHA Dataset

The MOCHA (Multi-Channel Articulatory) dataset, that forms the raw material for this project proposal, was recorded as part of an EPSRC project to evaluate articulatory data as feature set for training speaker-independent continuous ASR systems. The dataset includes 40 speakers of English each reading 460 TIMIT sentences (British version). The articulatory channels include EMA sensors directly attached to 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. The Laryngograph/EGG measures changes in the contact area of the vocal folds, providing pitch and voiced/voiceless information. EPG provides tongue-palate contact data at 62 normalised positions on the hard palate, defined by landmarks on the upper maxilla. EPG includes lateral tongue contact information which is missing from the EMA data. In addition, video footage of the mouth region is recorded. The dataset is already being used by a few speech technology researchers [Richardson et al, 2000][Frankel et al, 2000].

![Diagram of articulatory gestures and measurement instruments](image)

Figure 1. Placement of measurement instruments

Gestures

- Tongue body constrictions (extrinsic muscles)
- Tongue tip
- Tongue body shape (intrinsic muscles)
- Lip gestures
- EPG
- Velar port opening
- Pharyngeal constriction
- Larynx raising
- Vocal fold tension
- Glottal opening
- Laryngograph
- Subglottal pressure

Proceedings of the Institute of Acoustics

EMATools is a graphical user interface for analysing EMA, EPG and EGG and acoustic data simultaneously. The original code was written by Noel Nguyen [Nguyen, 2000] and has been subsequently modified by the author to read file formats based on NIST SPHERE and Edinburgh Speech Tools standards, and to make general improvements in usability, accuracy and display options. The GUI currently runs under Matlab which can be purchased for Apple Mac, PC and UNIX. As well as providing platform independence, Matlab is available in many speech production laboratories and permits users to modify the code easily to adapt it to their own requirements. EMATools is available free on the internet and is used currently to view datasets (including MOCHA) produced at the Edinburgh speech production facility3 ASR using MOCHA-TIMIT.

3.1 Baseline System

![Diagram of the baseline system used for all experiments](image)

The ASR training and testing was performed using a jack-knife procedure where test group 1 consists of 92 sentences numbered 1, 5, 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 [Woodland, 2000]. Acoustic features were mel-scaled cepstral coefficients (mfccs) 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 ltering 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 2nd 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. A phone bigram was trained using all 460 sentences. 46 monophone 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 jack-knife test set and found to be the same for both speakers.

Figure 3: Accuracy for baseline acoustic system using N mfcc coefficients + 1 energy coefficient + Δ + ΔΔ input features. Trials for N=12,14,16 and 18 mfccs are shown for the female speaker fsew0 and male speaker msak0 with 2-7 Gaussian mixture models.

Baseline results using MOCHA-TIMIT acoustic data are comparable with results for the original TIMIT. It should be noted that the best performance for fsew0 (65% accuracy) was with 14 mfccs and for msak0 (63.5% accuracy) 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.

3.2 Articulatory Features

In order to use the raw articulatory data as input to an ASR system the separate data channels must be combined; correlated components removed and the dimensionality reduced. As a first attempt, principal components analysis (PCA) is an obvious candidate process for achieving this.

3.2.1 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.
Proceedings of the Institute of Acoustics

3.2.2 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 [Wrench & Hardcastle, 2000].

3.2.3 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.

3.3 Articulatory Feature Vector Using Principal Components

The combined articulatory feature vector was built by applying principal components analysis to the 62 dimensional binary EPG data and saving the first 4 EPG principal components. These were concatenated with the voice energy value and the corresponding deltas and deltadeltas to create a 15 dimensional vector. This was in turn concatenated with 30, 36 or 42 principal components derived from applying PCA to the EMA data, creating an overall vector size of either 45, 51 or 57. PCA was applied once more in each case to diagonalise the covariance matrix without further reducing the dimensionality.

The best performance for speaker fsew0 is 55% accuracy, achieved using 30 EMA principal components (figure 4).

3.4 Articulatory Feature Vector Using Linear Discriminant Analysis

The following set of experiments were carried out on one female speaker. Principal components analysis combines the feature vectors, reduces their dimensionality and diagonalises the covariance matrix generated by the resulting vector. It is a convenient method but 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 (LDA) applied to carefully selected classes.

As a first attempt at using LDA, the 46 phones were selected as the class basis for discrimination. The combined articulatory feature vector was built by taking the 14 EMA channels +Δ + ΔΔ making 42 parameters in total and applying LDA based on the 46 phone classes arriving at 42 LDA components. The recognition accuracy achieved with this vector was 55% (Figure 5). This EMA-based feature vector can be enhanced by concatenating a measure of voicing energy derived from the laryngograph + Δ + ΔΔ and performing LDA on the combined vector. The resulting 45 dimensional vector provides an improvement in recognition accuracy of 8% giving 63% for the speaker fsewO (figure 5).

3.5 Combined Feature Vector

Previous research has shown that the combination of articulatory and acoustic features can result in an improved recognition accuracy compared with either the acoustic or articulatory features on their own [Wrench, 2000]. A previous paper by the author [Wrench & Richmond, 2000] showed that using

\[
\begin{align*}
&\text{no. Gaussian mixtures} \\
2 &\quad 3 &\quad 4 &\quad 5 &\quad 6 &\quad 7 \\
\text{Accuracy (\%)} &\quad 75 &\quad 70 &\quad 65 &\quad 60 &\quad 55 &\quad 50 \\
\text{EMA42 (LDA)} &\quad \text{(EMA42 (LDA) + Voice} &\quad \text{(EMA42 (LDA) + Voice}} &\quad \text{Acoustic45} &\quad \text{(EMA42 (LDA) + Voice} &\quad \text{Acoustic45)} &\quad \text{Acoustic45)} \text{51 (PCA)}
\end{align*}
\]

Figure 5: Mean accuracy results for fsewO. Combination of LDA-based articulatory feature vector with acoustic vector using principal components analysis to reduce the dimensionality. Compared with LDA based articulatory features and acoustic features on their own. 95% confidence limits are marked.

PCA derived articulatory features combined with the acoustic set and using PCA to perform dimension reduction, that recognition accuracy could be improved by 2%. Here we show that if the improved LDA based articulatory features are combined with the acoustic features, the improvement is a more impressive 6%, raising the accuracy to 71% (Figure 5). Note that PCA is still used in this case to combine the articulatory and acoustic data since the optimal combined output dimension is greater than can be generated using LDA on 46 classes. We would need to increase the number of classes to 52 to generate a combined vector of dimension 51.

Using the same data, jack-knifing procedure and underlying system a word recognition accuracy score was measured. The best accuracy using 45 mfcps was achieved with 6 Gaussian mixtures and the score was 59%. Using exactly the same setup with the above combination of acoustic and articulatory features the recognition accuracy increased to 68%. The improvement of 9% is significant given 95% confidence limits of around +/- 3%.
4. DISCUSSION

Our understanding of how to process articulatory data is still very primitive. The TIMIT dataset on which the MOCHA-TIMIT dataset is based was the basis of acoustic speech recognition studies 20 years ago. It took many years to come to a consensus on the best way to process the acoustic speech stream to optimise recognition accuracy. With articulatory data we are not dealing with just one new stream but many and subsets of these such as EMA and EPG differ in their characteristics. It is unlikely, therefore, that a few experiments by one or two researchers will find an optimal method of using this data in speech technology. Nevertheless, the experiments reported in this paper demonstrate that articulatory data at least has the potential to provide a route to a new generation of speech technology.

There is, however, a key step missing, without which an articulatory basis for speech technology is severely limited and that is the estimation of the articulatory data from the acoustics: The so-called acoustic to articulatory inversion problem. The improvements demonstrated in this paper can only be realised in a practical system that has a microphone or similar remote transducer as its interface. Connecting people to a collection of articulatory measurement instrumentation to perform speech recognition is not realistic. In the end these improvements in accuracy (or a significant fraction of them) must carry through to a system which estimates the articulatory data from the acoustics. At first glance it may seem unreasonable that any system derived from acoustic input can better the cepstral representation which has always shown its superiority despite many man-years of research focused on finding something better. However, with the MOCHA-TIMIT database it is now possible to take a step beyond abstract models of speech production such as the acoustic-tube model and apply supervised data-driven techniques to learn the relationship between acoustic output and the underlying constrained speech production system. The many-to-one mapping problem, which can be demonstrated by having more than one setting of an acoustic tube model produce the same acoustic output, may diminish when the dynamics of the articulatory system can be learnt. Collaborative research is already making headway on this crucial problem [Richmond, 2001].

The current experiments are limited to speaker-dependent recognition. Speaker-independent experiments will require an investigation of how best to normalise the articulatory measures across speakers. Some non-linear deformation of the geometric space in which the EMA sensors move is most likely in order to account for the differences in shape and size of the articulators.

On a more general note this database allows both speech scientists and technologists to examine the hidden articulatory values on which linguistic feature systems are based. The phone-based recognition task is chosen to demonstrate the power of the articulatory data because it is familiar. However, studies of this data may reveal an alternative unit upon which to base synthesis and recognition technologies. Indeed, the sequential unit based approach may give way to a parallel overlapping unit approach. This database offers an unprecedented opportunity to model the relationship between the acoustic output and the underlying articulatory system.

An articulatory feature space is constrained by biomechanical properties and therefore we would argue that has the potential not only to improve the accuracy when training an automatic speech recognition (ASR) system but also to form a basis for picking and merging units for synthesis or to expose idiosyncratic articulatory settings and dynamics for speaker recognition systems.

In summary we would encourage the widest possible examination and exploitation of this database in order that at least some of the many possible avenues of research can be pursued in the next 10 years.
5. CONCLUSIONS

We have presented results for a very standard speaker-dependent sentence-based phone recognition system, tuned for optimum performance using the acoustic MOCHA-TIMIT data. The results show that the recognition accuracy of the system based on purely articulatory data is comparable with the accuracy of an identical system based on purely acoustic data. The difference of 2% between 62.5% accuracy for the articulatory feature input and 65% for the accuracy using a standard acoustic input vector is significant but only just.

We have shown that articulatory data provides complementary information resulting in a significant improvement in recognition accuracy when articulatory and acoustic inputs are combined. The latest results using acoustic and articulatory features combined using linear discriminant analysis on phone-based classes provide a recognition accuracy of 71% - an improvement on the acoustic-only baseline of 6%.

There is still a great deal that can be done to enhance the articulatory feature selection and hence improve the accuracy of recognition using articulatory data. The choice of whole phonetic segments as classes is somewhat crude. Some sort of subphonetic class exhibiting more homogeneity may well provide further improvements. A whole new articulatory measurement (not currently included in MOCHA-TIMIT) revealing glottal spread is currently being investigated. This additional information is indicative of voicing intent and may reduce confusion between voiced and voiceless consonants. Should it prove useful, it will be added to future recordings for the expanding database.

It is our intention to make further significant improvements to the above system by refining the discriminant classes and adding new channels and we plan to work towards a practical system by using a multi-layer perceptron to estimate the articulatory data from the acoustics in the recognition phase.

The key to this research is the MOCHA database. The database is freely available for academic research. See http://www.cstr.ed.ac.uk/artic for release details.

ACKNOWLEDGEMENTS

This work was funded by the Engineering and Physical Sciences Research Council grant number GR/L78680

REFERENCES


Proceedings of the Institute of Acoustics


Proceedings of the Institute of Acoustics


