Automatic Speech Recognition Experiments with Articulatory Data

Esmeralda Uraga, Thomas Hain

Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, UK
{e.uraga,th}@dcs.shef.ac.uk

Abstract

In this paper we investigate the use of articulatory data for speech recognition. Recordings of the articulatory movements originate from the MOCHA corpus, a database which contains speech, EGG, EMA and EPG recordings. It was found that in a Hidden Markov Model (HMM) based recognition framework careful processing of these signals can yield significantly better performance than that obtained by decoding of the acoustic signals. We present detailed results on the processing of the signals and the associated performance of monophone and triphone systems. Experimental evidence shows that acoustic-signal-to-word mappings and articulatory-signal-to-word mappings are equally complex. However, for the latter, evidence of short-comings of standard HMM based modelling is visible and should be addressed in future systems.

Index Terms: articulatory representations, speech recognition.

1. Introduction

To find a robust and optimal speech representation for speech recognition has been a long standing objective in the research community. Recently, the interest in using articulatory features has been growing. Previous studies have found that articulatory features are more robust to noise than acoustic representations [1]. Our first goal towards the design of acoustic-to-articulatory mapping methods is to investigate whether articulatory representations contain sufficient information to be of use for speech recognition.

Standard front-ends in speech recognition systems are usually based on representations that are only moderately related to speech production mechanisms. In the source-filter model of speech production, the source represents the air flow at the vocal cords, and the resonances of the vocal tract are represented by the filter. This model has led to design methods of source-filter separation from the speech signal. Cepstral analysis and linear prediction are common source-filter separation methods. Despite their limitations to obtain proper representations of vocal tract resonances [2], they have been used to obtain most standard acoustic representations such as Mel Frequency Cepstral Coefficients (MFCC) [3] and Perceptual Linear Prediction coefficients (PLP) [4]. Although the main motivation for MFCC and PLP is to model perceptual behaviour.

Measurements of the location and movement of the articulators and the vocal cords should, in theory, contain sufficient information to reconstruct the speech signal and hence can be thought of as representation in very compact form. However, in practice the measurements are constrained to certain key components (for example the position of the tongue tip) rather than the complete speech production organ configuration. It is not known whether, even with correct modelling of parameter movement, optimal recognition performance can be achieved. Hence recognition errors will be caused by measurement techniques as well as parameter modelling. Experiments in this paper make use of recordings provided in the MOCHA corpus [5] which provides high quality recordings of the main parameters (see below).

Acoustic representations for speech recognition have been developed and optimised in the recent years [1]. In contrast, there is little experience in using articulatory representations of speech. It is very likely that similar efforts to those that led to the development of the currently popular acoustic representations based on MFCC or PLP coefficients may be necessary to obtain optimal representations of articulatory data.

In this paper, we compare the use of articulatory representations with the use of acoustic representations for the task of speech recognition with the objective of understanding the basic properties of underlying parameters governing speech generation. Contrary to previous experience, we found that the use of articulatory representations outperforms recognition with standard MFCC features. Analysis of the results suggests that further improvement should be achievable by improving the modelling of voicing and better handling of temporal variation.

The rest of the paper is organised as follows: a brief overview of the MOCHA corpus precedes an outline of our baseline recognition system. Then details on feature pre-processing of articulatory data are presented, followed by an analysis of experimental results. The final section provides a summary and conclusions.

2. The MOCHA corpus

The Multi-Channel Articulatory speech database (MOCHA) consists of parallel acoustic speech and articulatory data [5]. The articulatory data consists of recordings of electro-magnetic articulograph (EMA), electro-glottograph (EGG) and electro-palatograph (EPG) signals. The EMA data is a 14-channel signal including x and y coordinates of 7 pellets positioned at the soft palate, tongue dorsum, tongue blade, tongue tip, jaw, upper and lower lips. The EPG signals provide additional information by measuring binary tongue-palate contact data at 62 location-normalised positions on the hard palate. Finally, the EGG measures changes in the contact area of the vocal folds.

While the original MOCHA recordings covered 25 speakers, only data from 3 speakers are currently available in fully processed and annotated form. 460 sentences from each speaker were recorded. The corpus includes automatically-generated phone-level transcriptions and a keyword dictionary. The phonetic transcriptions make use of a phonetic alphabet of 46 symbols (44 En-

glish phones, silence and breath) with pronunciations selected to specifically cover the speaker's accent. For comparison with previous work [5] and to avoid problems related to inter-speaker variations due to vocal tract anatomical differences, experiments in this paper make use of data from only one British English female speaker (fsew0) which covers about 30 minutes of speech.

3. Baseline system

An acoustic representation based on 39-dimensional feature vectors was extracted from the speech signals provided in the fsew0 data set (MOCHA). These vectors consisted of 12 MFCC features plus acoustic energy (AcE), computed every 10ms over a 25ms window, as well as their first and second order derivatives. All systems presented in this paper are based on standard Hidden Markov models and N-gram models (where appropriate) and make use of standard HTK [6] for training and testing. The recognition systems constructed for this data are closely related to our baseline system for the TIMIT database. With the procedure outlined below the best system performance on the TIMIT core test set was an accuracy of 73.7% which is a very competitive performance for non-discriminatively trained systems. The acoustic training procedure involves maximum likelihood training of 3 state left to right monophone models from scratch. Models are trained using a standard HTK mix-up procedure [6]. Breath models are excluded from both training and testing and all breath segments are merged with silence. The phone time boundaries provided with the transcriptions are only used in the model initialisation phase. Triphone models are initialised from monophone models using 2 model re-estimation [6]. State clustering with phonetic decision trees and the mix-up procedure yield the final model sets.

Due to the small amount of data the complete fsew0 data set was split into five equally-sized subsets and five-fold cross-validation was used for training and testing. Recognition performance is computed as average over the five sets. Cepstral mean normalisation was used in all experiments in this paper. Table 1

Dim	Acoustic model	Phone LM	%PER
39	Monophone	TIMIT	35.7
39	Monophone	MOCHA	32.8
39	Triphone	MOCHA	30.0

Table 1: Phone error rate (%PER) for systems operating on acoustic features.

shows phone error rate (PER) results using phone bigram language models obtained from TIMIT or MOCHA. Even though TIMIT contains much more data the difference in accent is evident from improvement when using bigrams trained on MOCHA data. All acoustic models were trained using 4/5 of data (fsew0). The lowest phone error rate (30%) was obtained with a triphone system (which we refer as MFCC39).

4. Recognition using articulatory information

Speech recognition systems using articulator input operate in exactly the same way as those described in section 3. The systems only differ in terms of input feature representation. The clustering thresholds in phonetic decision tree state-tying are set to yield an approximately identical number of states for each system. This can lead to an increase in the number of parameters when using higher dimensional feature vectors. However, in all of the cases

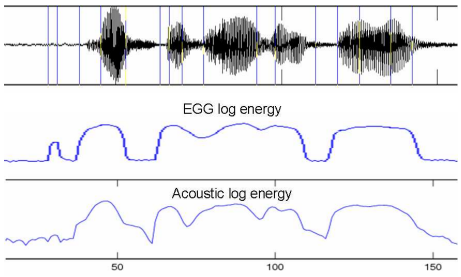

Figure 1: Sentence "This was easy for us". Speech waveform, energy contour of EGG signal and acoustic energy (AcE) contour.

observed, an increase of the number of parameters of the baseline system did not lead to significant performance improvement.

4.1. Articulatory data representation

The EMA data consists of x and y coordinate pairs of the moving pellets, recorded with a sampling rate of 500Hz. Closer inspection reveals mostly the presence of noise in frequencies above 50Hz (including a strong 50Hz power supply component). Hence a third order Butterworth filter with a cutoff frequency of 45Hz was applied to each recording and the signal was downsampled to 100Hz, identical to the MFCC frame rate. The EPG data comes in the form of 62 binary signals sampled at 200Hz. Assuming that the location and extent of touch of the main tongue body constitutes the essential information, means and variances in both x and y direction are computed on a per frame basis, thus yielding 4 parameters without further filtering. The EPG parameters are downsampled to 100Hz.

4.2. Energy components

The EGG recordings allow inclusion of reliable voicing information in the recognition process. Since larynx movement information is thought to be of little relevance for recognition the signal was filtered using a bandpass filter with lower and higher cut-off frequencies at 60Hz an 105Hz respectively. After filtering, the raw energy was computed every 10ms over a 25ms window. In addition, a signal corresponding to the raw pressure energy originating from the lungs would be required. Since such a signal is not available the raw acoustic energy, identical to that used in the acoustic system (see section 3), was used. Figure 1 shows both energy signals for a sample utterance. It is clear that the EGG energy signal approximates binary voicing information whereas acoustic energy retains information in unvoiced parts.

4.3. Phone recognition experiments

Experiments were conducted with different combinations of the articulatory features described above. Table 2 shows phone error rates for monophone and triphone models. All monophone models have exactly the same number of states, while in all triphone models this can only be achieved approximately. Note that within each experiment five different model sets with five slightly different numbers of parameters are trained. In all cases the number of parameters is a function of the feature vector dimension. The dimension is noted in the table, and includes first and second order derivatives for all static features. EMA data alone yields a performance substantially poorer than that shown in Table 1. The addition of EGG energy data yields a 16% relative reduction in PER, from which 8% is due to the application of the band pass filtering to the EGG signal. This can be further improved by adding

D	System	EMA	EPG	EGG	AcE	Mono	Tri
42	Art42	×				39.6	36.2
45	Art45a	×		×		33.9	30.3
45	Art45b	×			×	35.3	31.8
48	Art48	×		×	×	31.6	29.1
57	Art57	×	×	×		33.2	29.9
60	Art60	×	×	×	×	31.4	28.4

Table 2: % Phone error rates for systems operating on articulatory data from EMA, EPG and EGG sources as well as acoustic energy (AcE) for monophone (Mono) and triphone (Tri) systems.

acoustic energy information. Note that the gains in both cases are not additive, implying redundancy. This is not surprising as acoustic energy can be used as a good predictor for voicing information. However, as is also evident in results here, not all information is present. Furthermore, note that the recognition system based on articulatory information alone outperforms the results using acoustic features. In addition, information derived from EPG data yields another 1.5% absolute PER reduction. This result is likely to be lowered by a dramatic increase in feature vector size and hence, the number of system parameters estimated on a very limited training data size. A comparison of results for monophone and triphone systems show consistent behaviour throughout.

4.4. Parameter reduction techniques

In the experiments reported in [5], principal component analysis (PCA) is applied to the EMA data. In this work, PCA was found to degrade performance.

In additional experiments, we use knowledge about the vocal tract anatomy to derive a 9-dimensional representation of the articulatory data identical to the one used in [7] apart from two differences: here the acoustic energy component as used in the previously described system is added. Further a lip rounding feature is included, which is computed as the difference between lower lip and lower incisor positions in the horizontal direction. With this representation, the data dimensionality was reduced from 48 to 27. This data set was used to build a new speech recognition system (Art27). Using this alternative representation, the system performance was similar to that obtained with much higher dimensional feature vectors (Table 3).

System	%PER	%Voicing	%Place	%Manner
MFCC39	30.0	13.3	23.6	15.5
Art42	36.2	19.4	24.2	18.8
Art45a	30.3	15.2	21.6	16.6
Art45b	31.8	16.5	21.4	15.6
Art48	29.1	14.5	21.1	15.1
Art57	29.9	15.3	21.6	15.6
Art60	28.4	14.5	20.9	14.7
Art27	28.9	13.7	21.7	15.6

Table 3: % Phone error rate (%PER) for triphone systems. Error rate based on the clustering of phonemes in 3 phonetic dimensions: voicing, place and manner of articulation.

4.5. Analysis of errors

In order to identify possible patterns of errors, the confusion matrices were analysed in terms of 3 phonetic feature dimensions: voicing, place and manner of articulation [8]. Phonetic features used for voicing were voiced and unvoiced. Classes used for place of articulation include consonants and vowels. A consonant is classified as bilabial, labiodental, dental, alveolar, post-alveolar, palatal, velar or glottal. A vowel is classified by its height: low, mid or high, and by its backness: front, central or back. The features used for manner of articulation were stop, nasal, fricative, vowel, central approximant and lateral approximant. Silence was added as an additional state in each phonetic dimension.

For each phonetic dimension a deterministic mapping from phonemes to phonetic features was applied. Diphthongs were classified in the same category of their first vowel. Rows and columns in the confusion matrices corresponding to phonemes in the same phonetic feature category were clustered.

The results in Table 3 (Voicing) show that the level of recognition errors in the voicing dimension is relatively low. The acoustic system has the lowest PER (13.3%). This indicates that, contrary to expectations, the acoustic system discriminates better between voiced and unvoiced phonemes. Apparently the acoustic representation encodes sufficient voicing information in the acoustic energy and in the MFCC coefficients. In contrast, for pure articulatory representations the explicit voicing information in one feature (EGG energy) appears to be sub-optimal.

Table 4 shows an excerpt of the phone confusion matrix for the acoustic system. Note that albeit high, the confusion between plosives that differ by voicing information is not outstanding compared to confusion with other plosives. However, the equivalent matrix in the articulatory case (Table 5) shows more confusion errors between phonemes with same place of articulation (p/b, t/d, and k/g) and a clear distinction between plosives apart from these pairs.

	p	b	t	d	k	g	Del
p	273	26	16	3	27		19
b	22	240	3	7	3		17
t	11	1	668	42	23	3	89
d	4	2	69	324	2	3	77
k		1	19	1	465	15	32
g		4	4	9	51	101	12
Ins	26	15	108	60	44	11	

Table 4: Confusion matrix for plosives using acoustic features (raw counts). System MFCC39.

	p	b	t	d	k	g	Del
p	289	57	4				18
b	52	218	1				20
t	3	1	590	85	4	2	126
d	2	1	98	291	2		95
k	1		1	1	464	39	32
g			1		75	94	15
Ins	17	20	87	55	22	11	

Table 5: Confusion matrix for plosives using articulatory features (raw counts). System Art57.

In a similar way, the analysis of the results in Table 3 indicates that the best articulatory systems (Art48, Art60) outperform the acoustic system in the phonetic dimensions of manner and place of articulation. This becomes more evident when the percentages of correct recognition are compared in each phonetic dimension. The percentages in Table 6 show that the articulatory system Art27 outperform the acoustic system in correct recognition of consonants and vowels. The percentage of correct recognition is greater

than 90% for almost all the phonetic classes. In comparison with Art57 and Art60, the acoustic system shows better performance for vowel recognition. A possible explanation for the confusion errors between vowels is that phonemes or place of articulation features such as Low or Front that specify vowels, do not have well defined articulatory correlates [9]. Other confusion errors can be explained by examining the confusions between different consonants. They occur partially because their places of articulation are very close to one another. The only consonant classified as glottal is /h/. The poor recognition of /h/ may be because of its own nature. The articulatory configuration of /h/ can be that of any vowel. /h/ can be produced with a weakened voicing (between vowels) or no voicing (at the beginning of an utterance), but with the vocal cords vibrating. It may be possible that the consequent articulatory variability of these units was not appropriately modeled.

Place of Art.	MFCC39	Art27	Art57	Art60
Bilabial	88.4	97.6	96.2	96.7
Labiodental	80.8	91.9	94.4	94.1
Dental	61.9	88.2	91.8	92.2
Alveolar	95.2	98.3	96.4	96.5
PostAlveolar	80.9	93.0	80.8	78.1
Palatal	60.9	90.8	78.5	82.1
Velar	91.5	98.7	97.8	97.5
Glottal	83.1	83.2	69.8	68.1
Silence	100.0	100.0	100.0	100.0
High-Front	86.3	95.4	89.2	86.3
Mid-Front	82.8	93.7	82.6	81.1
Low-Front	87.5	94.5	84.1	85.6
Mid-Central	83.9	91.9	84.2	82.8
High-Back	82.0	92.3	83.4	84.8
Mid-Back	81.3	89.6	80.6	81.0
Low-Back	79.1	86.8	77.0	76.6

Table 6: Illustration of the patterns of correct recognition resulting from the clustering of phonemes per place of articulation in different confusion matrices.

5. Word recognition experiments

The objective in automatic speech recognition normally is not the recognition of phonemes but words. Phonemes are considered as an intermediate step that allow simple and effective discrimination in a lower dimensional space.

Word recognition experiments where conducted for both the acoustic front-end and the best articulatory front-ends. Table 7 shows word error rate results for both systems, using a word loop grammar and a vocabulary of 1817 words. Notably, the difference in performance between monophone and triphone system has increased to approximately 5% absolute. It is interesting to note that the performance differences between acoustic and articulatory representations are similar to those observed using phoneme error rate. Due to word constraints, the underlying phoneme error rate is considerably lower than that obtained in free phone recognition. Note however, that the reduction in PER for the Art60 system is 12% relative and thus, far more than the WER reduction.

6. Conclusions

Experiments with articulatory data have shown that performance better than that of associated acoustic systems can be obtained.

Dim	Front-end	WER Mono	WER Tri	PER Tri
39	MFCC39	33.3	28.0	15.4
27	Art27	-	27.8	15.2
57	Art57	-	27.7	15.4
60	Art60	30.5	26.0	13.5

Table 7: % Word error rate (%WER) and % Phone error rate (%PER) for monophone (Mono) and triphone (Tri) models using a word loop grammar as language model.

We have demonstrated that EMA, EPG, EGG and acoustic energy hold independent information that can be exploited in an efficient way. We have proposed a new articulatory representation derived from standard signal processing techniques and a knowledge-based approach for parameter reduction. One of the advantages of this compact representation is that it allows to build simpler speech recognition models without performance degradation. Experimental evidence suggests that better feature extraction techniques and feature reduction schemes, such as factor analysis and linear discriminant analysis (or improved versions thereof), will allow for further improvements. It is clear that long-term dependencies play an important role and are not addressed in many works on articulatory motivated acoustic modelling.

7. Acknowledgements

We would like to thank Alan Wrench and Simon King for their help and sharing of valuable experience with MOCHA. Esmeralda Uraga is supported by the Mexican government (CONACYT).

8. References

[1] Kirchhoff K. (1999), *Robust Speech Recognition Using Articulatory Information*, PhD Thesis, University of Bielefeld, Germany.

[2] Krstulovic S. (2001), *Speech Analysis with Production Constraints*, PhD Thesis, IDIAP, Switzerland.

[3] Davis S.B. and Mermelstein P. (1980), *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*. TASSP, 28(4):357–366.

[4] Hermansky H. (1990), *Perceptual Linear Predictive (PLP) analysis of speech*. JASA, 87(4):1738–1752.

[5] Wrench A. (2001), *A new resource for speech production modelling in speech technology*. Proceedings of the Workshop on Innovation in Speech Processing.

[6] Young S., Evermann G., Gales M., Hain T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V. and Woodland P. (2004), *The HTKBook*, Cambridge University, UK.

[7] Richardson M., Bilmes J. and Diorio C. (2003), *Hidden-articulator Markov models for speech recognition*, vol. 41, Speech Communication.

[8] Ladefoged, P. (2001), *A course in phonetics*. Hartcourt College Publishers, Fourth edition.

[9] Ladefoged, P. (2005), *Features and parameters for different purposes*. Annual Meeting of the Linguistic Society of America.

[10] Wrench A. and Richmond K. (2000), *Continuous Speech Recognition Using Articulatory Data*, Proc. ICSLP 2000.