A MULTI-CHANNEL/MULTI-SPEAKER ARTICULATORY DATABASE FOR CONTINUOUS SPEECH RECOGNITION RESEARCH

Alan A. Wrench

Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, UK
a.wrench@sls.qmced.ac.uk, http://sls.qmced.ac.uk

Abstract

The goal of this research is to improve the performance of a speaker-independent Automatic Speech Recognition (ASR) system by using directly measured articulatory parameters in the training phase. This paper examines the need for a multi-channel/multi-speaker articulatory database and describes the design of such a database and the processes involved in its creation.

1. Introduction

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. However, despite a call (Rose et al., 1994) by the ASR community for a combined articulatory/acoustic corpus, no database suitable for the task of speaker-independent continuous speech ASR training currently exists. This is mainly because, until recently, measurement equipment had not been developed to the stage where the recording of large multi-channel corpora was feasible. Hitherto, research in this area has been confined either to 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) or using restricted articulatory datasets (Papcun et al., 1992; Zlokarnik et al., 1995; Jung et al., 1996; Zacks, 1994; Zlokarnik, 1995; King & Wrench, 1999). Several research groups are predicting that speech production modelling will enhance the performance of ASR systems (McGowan & Faber, 1997). Studies supporting the theory of Articulatory

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 1-13.

Phonology (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.

1.1. *ASR based on directly measured articulatory data*

So far, the majority of attempts to develop articulatory feature based models for ASR have been implemented using acoustic data with feature values derived from linguistic rules. By comparison, there have been fewer experiments using small amounts of directly measured articulatory data (summarised in Table 1). Papcun et al. (1992), whose work is based on the Articulatory Phonology theory of Browman and Goldstein (1992), were the pioneers in this area of research. They used x-ray microbeam data to train a very small scale ASR system. The task of the system, based on a neural net, was to identify articulations of English stop consonants. Separate 2-hidden-layer perceptrons were trained to model the movement of the lower lip, tongue tip and tongue dorsum using 25 frames (~200ms) of bark scaled FFT bins. They reported gesture recognition scores, stopping short of providing phone identification scores.

None of the published tasks have been very realistic, with the most difficult task being speaker-dependent isolated syllable recognition. Zlokarnik (1995) 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 V_1CV_2 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. (1999) 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 central). 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 before comparisons between acoustic and articulatory feature vectors are made.

1.1.1. *Acoustic to articulatory mapping*

As well as recognition experiments, there have also been some studies focussing on the estimation of articulatory data from acoustic data (Zacks & Thomas, 1994; Hogden et al., 1996; Roweis, 1997; Richmond, 1999). This process forms an essential step towards a practical ASR system based on articulatory data. To date these experiments have been single speaker experiments. With the exception of Zachs & Thomas and Zlokarnik, who tested estimated articulatory data as input to an ASR system, the acoustic-to-articulatory results are assessed by measurement of r.m.s error and Pearson product-moment correlations making it difficult to evaluate their efficacy as input to an ASR system.

In Zlokarnik's experiment, the articulatory movements during the testing phase were estimated using a multilayer perceptron that performed an acoustic-to-articulatory mapping. Under these more realistic conditions, when articulatory measurements are only available during the training phase, the error rate could be reduced by a relative percentage of 18 to 25% (cf. 60% with directly measured data).

Table 1. Summary of ASR experiments using directly measured articulatory data

Authors	Measure-ment	Stimuli	Spkrs	Test / Train	Total dataset	Evaluation	Score
Papcun et al. (1992)	x-ray microbeam for tt, td, ll*	six consonant schwa sequences CəCəCəCəCə for /p, t, k, b, d, g/	3	Jack-knife test data a) one utterance b) one speaker c) one consonant	Total 18 utterances per speaker	Handsegmented gesture identification	85/90 gestures correctly identified on test a) and b) 68/90 gestures correctly identified on test c)
Zacks et al. (1994)	x-ray microbeam for tt, tb, td, ll*	5 words 'keyed', 'caid', 'cod', 'code', 'cooed' repeated 3 times	3	Jack-knife test data on 1 speaker	Total 15 utterances per speaker	Vowel identification	39/45 correct identification using estimates of articulatory data
Zlokarnik (1993, 1994)	Carstens EMA for li, ll, tt, tb, td* Also an estimate of the above data based on acoustic input.	165 nonsense words in the form V_1CV_2 where V ∈ /a,i,u,y/ C ∈ /b,g,d,p,k,t, l,m,n,v,f,s,sh,h/ embedded in a carrier phrase "Ich sage bV_1CV_2 bitte" repeated twice	2	Train on one repetition. Test on the other.	Total 330 utterances per speaker (3mins each)	Speaker-dependent task isolated word recognition	85.8% word error rate using filterbank 94.8% adding articulatory data 89% adding an estimate of articulatory data derived from filterbank data C_1 100% V 96% C_2 82%

ul = upper lip, ll = lower lip, li = lower incisor, tt = tongue tip, tb = tongue blade, tm = tongue middle (spaced between blade and dorsum), td = tongue dorsum.

Table 1 (cont.). Summary of ASR experiments using directly measured articulatory data

Authors	Measurement	Stimuli	Spkrs	Test / Train	Total dataset	Evaluation	Score
Jung et al. (1996)	x-ray microbeam for ul,ll,tt, tb,tm,td* and EGG	330 confusable nonsense words in the form CVC where V ∈ /a,i,u,æ,ə/ C ∈ /b,g,d,p,k,t/ Embedded in "Say a word1 of a word2 again"	1	Stage 1 used 50 CVCs to test and the remaining 390 to train Stage 2 used 25 of the 50 CVCs to train and the remainder to test	Total 420 utterances (840 words)	Stage 1: Gesture recognition Stage 2: Hand segmented Phone recognition	Gesture recognition not directly evaluated C_1 100% V 96% C_2 82%
Soquet et al. (1999)	Movetrack for ul, ll,* and chin EPG anterior, posterior and central coefficients & pressure	1536 C1V1C2V2 nonsense words where V ∈ /a,i,e,o,u,y/ C ∈ /p,t,k,b,d,g/ /f/,/s/,/ʃ/,/v/,/z/,/ʒ/, /m/,/n/,/j/,/ʀ/	1	Jack-knife in 4 diphonemically balanced sets	Total of 1536 utterances	Speaker-dependent word recognition	Acoustic only 44.6% Articulatory only 36.8% Acoustic plus Articulatory 83% on test c)
King et al. (1999)	Carstens EMA for ul,ll,li,tt, tb,td*	Up to 16 repetitions of 16 isolated words in the form CVC where V ∈ /a,i,e,o/ C ∈ /d,p/	1	184 randomly selected words for training and the remainder for testing	Total of 248 utterances	Speaker dependent isolated word (CVC) recognition	Improved from 66% for LPC coefficient input to 89% for articulatory input

ul = upper lip, ll = lower lip, li = lower incisor, tt = tongue tip, tb = tongue blade, tm = tongue middle (spaced between blade and dorsum), td = tongue dorsum.

2. Articulatory Database

2.1. *Background*

Directly measured multi-channel articulatory data in large enough quantities to train speaker-independent ASR systems is 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. Lack of a database has severely restricted the nature of the research in this area to date. The only large database with tongue movement data which has been made available for continuous speech recognition experiments is the Wisconsin x-ray microbeam database, which consists of 60+ 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. This is not enough continuous speech data to perform ASR training on a comparable basis to state-of-the-art ASR systems.

2.2. *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 Electomagnetic Articulograph system, a Laryngograph, an Electropalatograph (EPG) and a microphone. The Laryngograph and microphone signals are recorded as two channels directly onto computer through a Soundblaster card installed in a PC. Another two PC's are used to record the EMA and EPG data directly. The three systems are synchronised using serial port communication combined with signal post-processing. The time available in a given session is limited at about 2.5 hours due to EMA sensors becoming detached. In this time it is possible to record up to 460 sentences.

2.3. MOCHA Database

The MOCHA (Multi-CHannel Articulatory) database is evolving to provide a resource for training speaker-independent continuous ASR systems. The planned 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 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 maxilla. EPG includes lateral tongue contact information which is missing from the EMA data.

2.3.1. *Recording setup*

The recording facility consists of a sound-damped studio and adjacent control room. The recording engineer initiates recording by a single key press on the EMA PC. The recording software then sends a serial port signal to start the acoustic/laryngograph recording on a second PC. Once the acoustic recording is initiated a confirmation signal is sent back to the recording software which then generates a 1kHz 20ms tone in the recording studio before sending a serial port signal to start the EPG recording on a third PC and then starting the EMA recording. The acoustic data is recorded using an Audio-technica ATM10a microphone placed approximately 40cm in front of the speaker. Both the laryngographic and acoustic signals are recorded directly onto a 16bit Soundblaster card at 16kHz. The EPG is recorded at 200Hz and the EMA at 500Hz.

2.3.2. *Post-processing*

Post processing is carried out to synchronise the channels and correct for EMA head movement. In order to synchronise the acoustic signal with the EMA, the start of the 1kHz tone is detected by correlating each utterance with an extracted sample of the tone, then the duration of the tone is subtracted from the acoustic recording. The laryngographic and acoustic signals are synchronised by correlating the residual signals generated from inverse filtering each signal. The inverse filter coefficients are generated every 10ms by 10th and 18th order LPC for the laryngograph and acoustic signal, respectively.

Rotation and translation of the 2D EMA sensor data is performed to ensure that the two reference coils (one on the upper incisors and one on the bridge of the nose) are coincident across all frames for a given speaker. This removes any component of head movement from the EMA data. A further rotation is performed to align the bite plane with the x-axis and a translation sets the origin at the position of the upper incisor reference coil.

The EMA coils placed on the tongue and soft palate are liable to become detached during the recording session. It is often possible to observe a coil in the process of becoming detached and to consolidate the bond before it falls off completely. In cases where the coil becomes completely detatched, it is replaced as close to the original location as possible. If necessary the EMA data can be post-processed to correct for discrepancies in coil placement. This is done by calculating the mean and variance of the x- and y-coordinates of the coil for each sentence and plotting them. If there is an observable discontinuity between the values before and after the coil is replaced then an offset and gain are applied to the coil data for the sentence after the coil is replaced in order to equalise the mean and variance. Offset and gain values are estimated separately for the x- and the y-coordinates. A coil is usually only replaced once during a session as the build-up of adhesive prevents satisfactory adhesion if subsequent replacement is attempted.

2.3.3. *Subjects*

The recording process requires a lot of co-operation from each subject during preparation and recording, particularly in the attachment of sensors, and consequently the subjects have to be willing and committed. To ensure these characteristics, subjects are mainly recruited from the student body, academic and support staff within the institution and from speech and language therapists who have connections with the institution. Candidates are screened for their ability to tolerate touching the soft palate and to obtain a Laryngograph trace. No attempt has been made to exclude subjects on the basis of dialect.

2.4. *Labelling*

Although, the presence of articulatory data may in the future allow automatic feature or gesture labelling, the initial labelling objective is to provide a phonetic labelling and a phonemic dictionary as is provided with TIMIT.

The labelling procedure currently consists of forced alignment of a phone sequence generated from a "Keyword" dictionary and word level transcription using flat-start monophone models.

2.4.1. *Keyword pronunciation dictionary*

The variety of dialects presents a challenge to the labelling procedure for the database. Pronunciation lexicons for use in speech synthesis and recognition are readily available for General American and Received Pronunciation (RP). Although rules can be generated to convert a standard lexicon (e.g. RP) to a target dialect semi-automatically (Fitt, 1997) the rules have to be substantially rewritten for each new accent. A modified solution is described by Fitt (Fitt & Isard, 1998; Fitt, 1999), based on Wells' keyword system (Wells, 1982).

> "Wells describes the vowels occurring in different accents in terms of keywords, so rather than saying that 'pool' contains the phoneme /u/ in RP and /ʉ/ in Scottish accents, he simply says that the word contains the GOOSE vowel."

Fitt's work is intended to generate dialect-specific lexica for synthesis, but can be used to provide a basis for pronunciation dictionaries for ASR as well.

For the task at hand a keyword pronunciation dictionary was generated for the 460 TIMIT sentences. Separate lexical rules for conversion of keyword symbols to phonemes have so far been generated for General American, Southern American, Welsh, Scottish, Northern English and Southern English. The postlexical rules which are applied to the dictionary in order to create phonemic sentence transcriptions are designed to cope with rhotic and non-rhotic dialects. Since phonetic realisation is speaker-dependent and therefore difficult to predict, no lexical or postlexical rules are currently applied at this level.

Figure 1 shows how a phonemic transcription might change for a given utterance according to dialect.

<table>
<tr><td>Keywords</td><td>dh @ | p r ow l @r r | w our r | @ | s k ii | m ah s k | f @r r | d i s g ae z</td></tr>
<tr><td>S. England</td><td>ð ə | p r aʊ l ə |w ɔ r | ə | s k i | m ɑ s k | f ə | d ɪ s g aɪ z</td></tr>
<tr><td>SW. England</td><td>ð ə | p r aʊ l ə r | w ɔ r | ə | s k i | m æ s k | f ə r | d ɪ s g aɪ z</td></tr>
<tr><td>Scotland</td><td>ð ə | p r aʊ l ə r | w əʊ r | ə | s k i | m æ s k | f ə r | d ɪ s g aɪ z</td></tr>
</table>

Figure 1. Dialect-specific phonemic transcriptions for the sentence "The prowler wore a ski mask for disguise."

3. Discussion

The baseline ASR system is currently being tuned using the Entropic HTK V2.1 system trained and tested on a single female speaker from the MOCHA database to achieve state-of-the-art phone recognition rates. The forced alignment method produces sufficiently accurate segmentation and labelling for the baseline ASR system to achieve phone recognition scores of 68%. However, a casual study of the label files reveals a significant number of transcription errors (approximately 2 per sentence) due to pronunciation variants not catered for in the single pronunciation dictionary (e.g. learned : /l ɜ r n d/ or /l ɜ r n ɪ d/); reading errors (e.g. 'the' for 'a') and co-articulatory processes. A principled method of improving the automatic transcription generation process is being sought.

4. Acknowledgements

This work was funded by the Engineering and Physical Sciences Research Council grant number GR/L78680. Thanks to Sue Fitt for helpful discussions with regard to preparing the keyword-based transcriptions.

5. References

Browman, C. & Goldstein, L. (1992). Articulatory phonology: An overview, *Phonetica* **49**, 155-180.

Carreira-Perpinan, M.A. & Renals, S. (1998). Dimensionality reduction of electropalatographic data using latent variable models. *Speech Communication* **26**(4), 259-282.

Deng, L. & Erler, K. (1992). Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units. *J. Acoust. Soc. Am.* **92**(6), 3058-3067.

Deng, L. & Sun, D. (1994). A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *J. Acoust. Soc. Am.* **95**(5), 2702-2719.

Fitt, S. (1997). The generation of regional pronunciations of English for speech synthesis. *Proc. 5th Conf. on Speech Comm. and Techn. (Eurospeech'97),* 2447-2450.

Fitt, S. & Isard, S. (1998). Representing the environments for phonological processes in an accent-independent lexicon for synthesis of English. *Proc. of the Int. Conf. Spoken Lang. Processing (ICSLP'98).*

Fitt, S. (1999). The treatment of vowels preceding 'R' in a keyword lexicon of English. *Proc. Int. Conf. of Phonetic Sciences (ICPhS'99)*, 2299-2302.

Hogden, J., Lofquist, A., Gracco, V., Zlokarnik, I., Rubin, P. & Saltzman, E. (1996). Accurate recovery of artticulator positions from acoustics: New conclusions based on human data. *J. Acoust. Soc. Am*. **100**(3), 1819-1834.

Jung, T-P., Krishnamurthy, A.K., Ahalt, S.C., Beckman, M.E. & Lee, S-H. (1996). Deriving gestural scores from articulator-movement records using weighted temporal decomposition. *IEEE Trans. Speech and Audio Processing* **4**(1), 2-18.

King, S. & Wrench, A. (1999). Dynamical system modelling of articulator movement. *Proc. Int. Conf. of Phonetic Sciences (ICPhS'99),* 2259-2262.

Kirchhoff, K. (1996). Syllable-level desynchronisation of phonetic fatures for speech recognition. *Proc. Int. Conf. on Spoken Lang. Processing (ICSLP'99),* 2274-2276.

Larar, J.N., Schroeter, J. & Sondhi, M.M. (1988). Vector quantisation of the articulatory space. *IEEE Trans. Acoust. Speech Signal Processing* **36**, 1812-1818.

McGowan, R.S. & Faber, A. (1997). Speech production parameters for speech recognition. *J. Acoust. Soc. Am.* **101**(1), 28.

Papcun, G., Hochberg, J., Thomas, T.R., Larouche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.* **92**, 688-700.

Ramsay, G. (1998). Stochastic calculus, non-linear filtering, and the internal model principle: Implications for articulatory speech recognition. *Proc. Int. Conf. on Spoken Lang. Proc. (ICSLP'98),* 2987–2990.

Randolph, M.A. (1994). Speech analysis based on articulatory behaviour. *J. Acoust. Soc. Am.* **95**(5), 2818 (A).

Richards, H., Mason, J.S., Hunt, M.J., & Bridle, J.S. (1995). Deriving articulatory representations of speech. *Proc. European Conf. on Speech Comm. and Techn. (Eurospeech'95),* 761-764.

Richmond, K. (1999). Estimating velum height from acoustics during continuous speech. *Proc. European Conf. on Speech Comm. and Techn. (Eurospeech'99),* 761-764.

Rose, R.C., Schroeter, J. & Sondhi, M.M. (1994). An investigation of the potential role of speech production models in automatic speech recognition. *Proc. Int. Conf. on Spoken Lang. Processing (ICSLP'94),* Vol. 2, 575-578.

Roweis, S. & Alwan, A. (1997). Towards articulatory speech recognition. *Proc. European Conf. on Speech Comm. and Techn. (Eurospeech'97),* 1227-1230.

Schmidbauer, O., Casacuberta, F., Castro, M.J., Hegerl, G., Hoge, H., Sanchez, J.A. & Zlokarnik, I. (1993). Articulatory representation and speech technology. *Language and Speech* **36**, 331-351.

Schroeter, J. & Sondhi, M.M. (1992). Speech coding based on physiological models of speech production. In: S. Furui and M.M. Sondhi (eds.), *Advances in Speech Signal Processing*, 231-268.

Soquet, A., Saerens, M., & Lecuit, V., (1999). Complementary cues for speech recognition. *Proc Int. Conf. of Phonetic. Sciences (ICPhS'99),* 1645-1648.

Wells, J.C. (1982). *Accents of English.* Cambridge: Cambridge University Press.

Zacks, J. & Thomas, T. (1994). A new neural network for articulatory speech recognition and its application to vowel identification. *Computer Speech and Language* **8**, 189-209.

Zlokarnik, I., Hogden J., Nix D. & Papcun G., (1995). Using Articulatory measurements in automatic speech recognition and in speech displays for hearing impaired (abstract). *ACCOR Workshop on Articulatory Databases*, Munich.

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