Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data

Claudia Canevari, Leonardo Badino, Luciano Fadiga, Giorgio Metta

RBCS, Istituto Italiano di Tecnologia, Genova, Italy

Abstract

We test an hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) phone recognition system that uses measured articulatory features as additional observations on two English corpora and an Italian corpus. The three corpora contain simultaneous recordings of speech acoustics and EMA (Electromagnetic Articulograph) data. We show that the additional articulatory features reconstructed from speech acoustics through an Acoustic-to-Articulatory Mapping, always produce a phone error reduction, with the exception of one single case where, however, the reconstruction accuracy of the articulatory features is significantly lower than in all other cases. Error analysis shows that in all corpora the articulatory features positively affect the discrimination of almost all phonemes although some phonemic categories are clearly more affected than others.

Index Terms: Acoustic-to-Articulatory Mapping, Electromagnetic articulograph, EMA, Deep Neural Networks, phone recognition

1. Introduction

Many phenomena observed in speech, such as, e.g., coarticulation effects, can be easily and compactly described in terms of vocal tract gestures but not on purely acoustic terms. That has been a strong motivation to use speech production knowledge for ASR [12]. When measured articulatory data are used the articulatory information can be incorporated in an ASR system by appending measured articulatory features to the standard observation feature vectors (e.g., MFCC vectors). Although such approach only exploits part of the potential utility of the articulatory data (e.g., it ignores their potential benefits for modeling speech dynamics), as opposed to strategies that explicitly model the cause effect relation between articulatory gestures and acoustics (see, e.g. [**?**]), it is one of the few strategies where articulatory features (AFs, either measured or "linguistic", i.e., extracted through phonetic rules) produced phone (e.g., [1, 2]) and word (e.g., [5, 7, 8]) recognition error reductions, especially in noisy speech conditions. In [1, 2] the use of measured articulatory data produced up to a 10.1% relative phone error rate reduction on the MOCHA-TIMIT dataset in clean speech conditions. In the present paper we assess whether similar results also apply to other datasets and languages and identify the phonemic categories whose discrimination is most positively affected by the use of AFs.

Like in [1] we used a Deep Neural Network - Hidden Markov Model (DNN-HMM) phone recognizer where Deep Neural Networks [11] are both used to estimate the phone posterior probabilities and to carry out the Acoustic-to-Articulatory Mapping (AAM) which allows to recover the AFs from speech acoustics. Recovering AFs is necessary in realistic scenarios where articulatory data are only available during training. The recovered AFs do not actually add any new information to the observation feature set, but are the result of a speech-production driven transformation of the acoustic domain that may result in an improved acoustic modeling.

The use of recovered AFs can be successful if their recovering is sufficiently accurate (i.e., if the AAM is good enough) and the method to estimate phone posteriors is able to exploit the transformed information provided by the recovered AFs. The reconstructed AFs can be (slightly) more effective (for phone recognition) if we distinguish between critical and non-critical articulators when performing AAM [3], and, we expect, if we achieve a better AF reconstruction, e.g., by applying some dynamic constraints on the reconstruction [16], or by applying methods that are able to handle the non-uniqueness of the AAM problem ([17, 20]). However, the fact that some studies showed a non-utility of the reconstructed AFs (e.g., [21]) whereas our DNN-HMM-based phone recognizer benefited from them [1] is most probably not only due to an improved AAM but also to the use of DNNs rather than mixtures of Gaussians for the computation of the observation probabilities.

In [1] we experimented with several AAM strategies but only on one dataset, specifically the msak0 voice of the MOCHA-TIMIT dataset [21]. An important question that follows that study is whether its results also apply to other datasets and languages. In the present paper we experiment with two additional corpora, the mngu0 corpus [19] and the "Lecce corpus" [9]. Like the MOCHA-TIMIT msak0 dataset, the mngu0 dataset was recorded from a native British English male speaker, with the main differences being a much larger collection of data and more accurate EMA data recording. The "Lecce corpus" was recorded from 9 native Italian speakers (but we could only use the data from 5 female speakers). It contains fewer utterances per voice w.r.t. the other datasets and each utterance is a single-word utterance. A cross-corpus and cross-linguistic evaluation not only allows us to assess whether the utility of the AFs stands across datasets and languages but also to identify "where and when" the AFs are more relevant. For example, the relevance of the AFs for phone recognition may depend on the training datasets (as it happened to be in a binary plosive consonant classification task, [4]) and on the phonemic category independently of the language.

2. Deep Neural Networks

A DNN-HMM ASR system is an HMM system where the observation probabilities are computed by means of a Deep Neural Network. The DNN-HMM framework have be recently shown to be very powerful for phone and speech recognition [14, 6]. The DNN computes the phone posterior probabilities (given the acoustic evidence) from which the observation probabilities can be easily computed.

In their standard formulation DNNs are feed-forward neural networks whose parameters are first pre-trained using unsuper-

vised training of Deep Belief Networks ([11]) and subsequently fine-tuned using, e.g., back-propagation. DNNs can be seen as an improved version of Feed-forward Neural Networks that exploits the knowledge of the statistical properties of the input domain (i.e., $P(X)$) to effectively guide the search for input-output relations (i.e., $P(Y|X)$).

The DNN training is carried out as follows. First a Deep Belief Network (DBN) is trained in an unsupervised fashion. Subsequently the DBN is transformed into a deep neural net by converting the stochastic activation function of each node into a deterministic function. If the DNN is used to perform regression or classification an output layer is added on the top. If the DNN is used for regression the output unit activation function is a linear regressor with linear basis functions while when it is used for classification the output unit activation function is a softmax function. Finally supervised fine-tuning of the parameters is applied.

A DBN is a hybrid probabilistic graphical model that can be trained by approximating it to a stack of Restricted Boltzmann Machines (RBMs). An RBM is an undirected graphical model with a layer of visible nodes ($\mathbf{v}$) and a layer of hidden nodes ($\mathbf{h}$) with intra-layer connections and without any within-layer connection.

The joint probability of an RBM is:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} exp(-E(\mathbf{v}, \mathbf{h})) \quad (1)$$

where Z is the partition function and the energy function $E(\mathbf{v}, \mathbf{h})$ for an RBM with both binary visible and hidden variables is:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i,j} v_i W_{ij} h_j - \sum_i b_i v_i - \sum_j c_j h_j \quad (2)$$

where W_{ij} are the connection weights and b_i and c_j are the biases on the visible and hidden nodes respectively.

Since there are no within-layer connections the probabilities $P(\mathbf{v}|\mathbf{h})$ and $P(\mathbf{h}|\mathbf{v})$ factorize and are given by:

$$P(v_i = 1|\mathbf{h}) = sigmoid(\sum_j W_{ij} h_j + b_i) \quad (3)$$

$$P(h_i = 1|\mathbf{v}) = sigmoid(\sum_i W_{ij} v_i + c_j) \quad (4)$$

The unsupervised learning of the parameters is performed by maximizing the $log(P(\mathbf{v})) = log(\sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h}))$. The gradient update rule for a parameter θ_k is :

$$\Delta\theta_k \propto < \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta_k} >_{data} - < \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta_k} >_{model} \quad (5)$$

where $< ... >_{data}$ stands for expected value under the empirical distribution and $< ... >_{model}$ for expected value under the model distribution. The latter can be computed by running block Gibbs sampling where $P(\mathbf{h}|\mathbf{v})$ and $P(\mathbf{v}|\mathbf{h})$ are sampled. Rather that running Gibbs sampling until equilibrium we can still effectively train RBMs by using contrastive divergence [10] where the Gibbs sampler can run for just one step.

RBMs with Gaussian distributed visible (or hidden) variables can be also trained by applying simple changes to some of the equations above.

A DBN can be trained by using layer-wise training where the output (i.e., the values of the hidden nodes) of a trained RBM is used as input for the RBM above. Then unsupervised parameter fine-tuning can be applied where the DBN is considered as a whole deep architecture.

Like in [1, 2] we used DNNs for both Acoustic-to-Articulatory Mapping and phone posterior estimation.

3. Corpora

We used three different datasets (see Table 1), all consisting of simultaneous recordings of speech and Electromagnetic Articulographic (EMA) data (plus other types of articulatory data that we ignored).

The first dataset is the MOCHA-TIMIT corpus ([22, 23]) which includes 460 British-English utterances for each of the two speakers, one male (msak0) and one female (fsew0). The second one is the mngu0 corpus described in ([19]). It consists of 1354 British-English sentences uttered by a male speaker. The latter dataset is the Italian Lecce corpus in ([9]) which consists of single-word utterances recorded from 9 Italian speakers. The dataset lexicon covers 73 different word types, either pronounced with a declarative intonation or a question intonation, and 65 pseudoword types. Each speaker was required to read out each type on average 3 times.

EMA data consist of the x and y positions of upper incisor (UI) (except for the mngu0 corpus), lower incisor (LI), upper lip (UL), lower lip (LL), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD).

We carried out our experiments on the msak0 voice of the MOCHA-TIMIT dataset, on the first 5 female subjects of the Italian Lecce corpus and on the all mngu0 dataset.

dataset	spk	utterances/words	phones	articulators
MOCHA-TIMIT	2	460	44	7
mngu0	1	1354	49	6
Lecce	6	642 (467)	42	7

Table 1: *Number of (i) speakers, (ii) recorded sentences or words, (iii) phones in each phone set and (iv) tracked articulators for the three datasets. In parenthesis the number of words pronounced only by speaker 2 in the Italian Lecce corpus.*

4. Experimental setup

From each dataset we used 60 mel-filtered spectral coefficients (MFSCs) as acoustic input for the AAM and as observations in the DNN-HMM phone recognition system. Contrary to [1] we used as acoustic observations MFSCs rather than MFCCs as it turned out that they produced slightly better results (in the baseline, i.e., acoustic observations only). This is in agreement with previous work in speech recognition based on DNNs (see, e.g., [15]). Concerning the articulatory data, we used 42 (36 in the mngu0 dataset) articulatory features (AFs) consisting of the x and y trajectories, plus their first and second derivatives, of the 7 (6 in mngu0) articulatory positions listed in section 3. Note that upper incisors exhibit very small variations in all phones and are used for head-movement corrections. In the mngu0 dataset they are not used as all other AFs since the provided articulatory data were already corrected. In the other 2 datasets upper incisors are reconstructed from speech acoustics as the other articulators.

Training and testing sets were created as described in ([21]) and in ([19]) for the msak0 dataset of the MOCHA-TIMIT and mngu0 dataset respectively. Concerning the Italian Lecce corpus for each of the 5 subjects we used two tokens (or at least one token in the case of subject 2) of each word type and pseu-

doword type in the training set and the remaining data in the testing set. As a consequence each word type occurred both in the training and testing dataset. That was necessary to have enough data to train the DNNs performing AAM.

4.1. Acoustic-to-articulatory mapping

The acoustic-to-articulatory mapping was performed by a 3-hidden layer DNN with 300 nodes per each hidden layer as in the simplest DNN configuration of ([1]). The input units of the corresponding DNN were Gaussian-distributed while all hidden units were binary. The input consisted of 5 acoustic feature vectors (60x5) or 1 acoustic feature vector (60x1) and the output was the vector of 42 (36 in mngu0) AFs corresponding to the frame on which the acoustic input is centered.

4.2. Phone recognizer

We used 3 states per phone. The state boundaries were computed using the HInit, HRest and HERest functions of the HTK ([24]).

The state posteriors were computed by a 3-hidden layer DNN, with 9 vectors of MFSCs (60x9 MFSCs) and the corresponding 9 vectors of AFs (42x9 or 36x9 AFs), when AFs were used as input. Each hidden layer had 1500 nodes while the output layer had 132, 147 or 126 units in msak0, mngu0 and Italian dataset respectively (44 and 49 British-English phonemes x 3 states or 42 Italian phonemes x 3 states).

In order to estimate the phone sequence for each test utterance we first computed the phone unigrams and bigrams and the state bigrams on the training data of each dataset for each split using the CMU toolkit ([25]). Then the state posteriors (not divided by the state priors) plus phone unigrams and bigrams and state bigrams were fed into a Viterbi decoder.

5. Results

5.1. Articulatory reconstruction

The AF reconstruction was evaluated using the Root-mean-square error (RMSE) and the Pearson product moment correlation (r). In table 2 the three datasets are compared on the reconstruction of the 36 articulatory features. The trajectories, velocities and accelerations of the upper incisors in the msak0 and Lecce dataset were excluded in order to have an unbiased comparison.

The articulatory reconstruction in the mngu0 dataset outperforms that in the msak0 dataset, especially when using 5 acoustic feature vectors in AAM as input. It confirms that the mngu0 corpus is a useful and good resource of acoustic and articulatory data, at least for the AAM problem, as reported in [18]. In fact the mngu0 dataset is claimed to provide a phonetically various and large set of data with very reliable articulatory measurements ([19]). While in MOCHA-TIMIT some articulatory inconsistencies are well documented mostly for fsew0 dataset, during the recording of mngu0 corpus no displacement of coils took place ([18, 19]).

Note that contrary to [18] here the mngu0 corpus was compared with the msak0 dataset rather than the fsew0 dataset.

Concerning the Italian Lecce dataset, the RMSE and r values reported in table 2 were averaged over four speakers, without considering speaker 3. Removing speaker 3 was motivated by the fact that the reconstruction of the AFs was very poor. In order to understand why the reconstruction was significantly worse than for all the other speakers we computed the correlation between actual AFs that belonged to the same phoneme in the same word type. Such correlation can be seen as a kind of intra-speaker articulatory coherence. The coherence of speaker 3 turned out to be significantly smaller than in all other speakers (Figure 1). This lack of articulatory coherence might be due a displacement of the coils during recording.

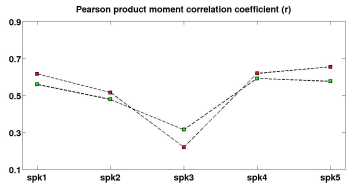

Figure 1: *Average correlation between actual articulatory features of the same speaker extracted from different instances of the same wordtype and pseudowordtype (green squares) and average correlation between actual and reconstructed articulatory features for each speaker (red squares) in the Italian Lecce dataset.*

Dataset	MFSCs vector input in AMM			
	5 vectors		1 vector	
	RMSE	r	RMSE	r
msak0	0.650	0.750	0.679	0.726
mngu0	0.542	0.837	0.660	0.744
Lecce	0.735	0.648	0.791	0.561

Table 2: *Articulatory reconstruction results in terms of Root Mean Square Error (RMSE) and Pearson product moment correlation (r) for each dataset. The input of AAM consists of 5 or 1 acoustic feature vectors. Values are averaged on the 5 splits in msak0 dataset and on the 4 subjects (subjects 3 excluded) in the Italian Lecce dataset. For an unbiased comparison the upper incisor trajectories, velocities and accelerations of the Lecce corpus and the msak0 dataset were excluded.*

5.2. Phone recognition

Table 3 shows the frame-wise classification accuracy (FwCa) and the phone error rate (PER) for the DNN-HMM phone recognition system using different types of observation sets in the three datasets.

In msak0 and in mngu0 the recovered AFs always improve phone recognition. The PER reduction ranges from 3.1% to 9.8% w.r.t. the acoustic baseline. A perfect articulatory reconstruction would lead to a 25% and a 20.6% PER reduction respectively. Note that the two British-English datasets use slightly different phone sets (different number of allophones for the same phoneme). For this reason a more fair comparison might be carried oot after conversion of one phone set into the other one.

In the Lecce corpus the PER using reconstructed AFs combined with acoustic ones is worse than that using only acoustics, while not considering speaker 3 the PER reduction provided by the articulatory features recovered from 5 acoustic feature vectors, turns out to be about 2% w.r.t. the acoustic baseline. A perfect articulatory reconstruction would lead to 38.6% (47.5% without speaker 3) w.r.t. the acoustic baseline.

It is important to point out that the PER reduction produced by the AFs recovered from 5 acoustic feature vectors might be

Feature set	msak0		mngu0		Lecce	
	FwCA %	PER	FwCA %	PER	FwCA %	PER
MFSCs	68.0	30.0	83.6	13.4	84.3 (84.6)	12.4 (12.5)
MFSCs + actual AFs	74.9	22.5	87.5	10.7	88.4 (89.3)	7.6 (6.6)
MFSCs + rec AFs 5-1	70.8	28.2	85.5	12.1	84.7 (85.7)	13.6 (12.2)
MFSCs + rec AFs 1-1	69.9	29.1	85.0	12.9	84.7 (85.5)	14.6 (14.4)

Table 3: *Frame-wise phone classification accuracy (FwCa) and phone error rate (PER) for each dataset using MFSCs only, MFSCs and actual AFs, MFSCs and AFs reconstructed from 5 acoustic feature vectors (rec AFs 5-1) or 1 acoustic feature vector (rec AFs 1-1). Values are averaged over the 5 splits in msak0 dataset and over the 5 subjects in the Italian Lecce dataset. In parenthesis values without considering subject 3.*

due to the implicit use of a larger acoustic context. In fact we might actually implicitly observe information from more than 9 acoustic frames (specifically 2 frames due to the first AF vector + 2 frames due to the last AF vector). In order to find out if that was the case we perfomed AAM by only using one vector of MFSCs as input. It turned out that the phone recognizer that uses 9 MFSC vectors + 9 vectors of AFs, each one recovered from the corresponding MFSC vector, outperformed the 9-frame acoustic baseline in mngu0 and msak0 datasets while in Lecce dataset the PER was worse. This last controversial result is most probably due to the fact that not enough there was not enough data to train a DNN to perform AAM with just one acoustic feature vector as input. That is confirmed by the poor articulatory reconstruction results showed in table 2.

5.3. Analysis of error

The PER was computed through the evaluation of the Levenshtein distance that is the minimum number of phone edits necessary to change the phone sequence estimated by the Viterbi decoder into the real one. Three types of possible phone edits/errors are considered: substitution, deletion and insertion.

Table 4 separately the relative error reduction produced by the AFs for each phone edit type due. For all datasets the AFs, both the actual ones and the ones recovered from 5 acoustic feature vectors, seem to affect phone substitution, while their contribution to phone deletion and insertion is less clear.

dataset	sub	del	ins
msak0	6.5 (33.3)	3.9 (5)	8.4 (32.9)
mngu0	17.1 (37.2)	1.1 (-49.4)	12.2 (56.1)
Lecce	4.7 (39.6)	17.8 (20.5)	-0.7 (53.7)

Table 4: *Relative reduction (%) of phone substitution (sub), deletion (del), insertion (ins) rate provided by reconstructed (from 5 acoustic vectors) and actual (values in parenthesis) AFs, w.r.t. the acoustic baseline for each dataset. Values are averaged over the 5 splits in msak0 dataset and over the 4 subjects (subject 3 was not included) in the Italian Lecce dataset.*

In order to better understand which phonemic categories are more positively affected by the articulatory features we ranked all phonemes in descending order from those phonemes that mostly benefited from the use of AFs to those that do not take any advantages in terms of frame-wise classification and phone substitution error (Figure 2). We chose the SAMPA standard to identify unambiguously all phoneme symbols in all three datasets.

In the two British-English datasets recovered articulatory information acts more positively in terms of frame-wise classification accuracy and phone substitution error reduction on

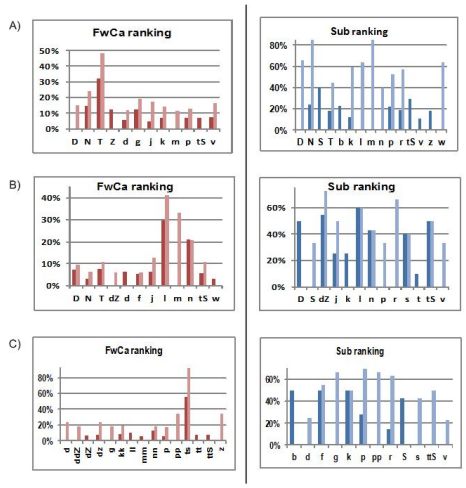

Figure 2: *Phone ranking in terms of FwCa improvement (left panel) and phone substitution error reduction (right panel) w.r.t. the acoustic baseline in msak0 dataset (A), mngu0 dataset (B) and Lecce dataset (C). For each dataset the two histograms show not more than the first 10 phonemes that mainly benefit from using of AFs: the actual ones in pink and light blue columns and the ones reconstructed from 5 acoustic feature vectors in red and blue columns. On the y-axis of the two histograms the values of FwCa relative improvement and phone substitution error reduction respectively. On the x-axis the phonemes are in alphabetic order.*

some fricative and nasal phonemes and on the two affricates: 'S', 'T', 'Z', 'N', 'tS' and 'D', 'T', 's', 'N', 'l', 'n', 'tS', 'dZ' in msak0 and mngu0 datasets respectively. Regarding the Italian Lecce dataset the single affricate 'ts' and some geminates as the nasals, 'mm', 'nn', 'll', 'kk', the affricate, 'ttS' and the plosive, 'tt', mostly benefit from using recovered articulatory information in terms of frame-wise classification accuracy, while some single plosives, as 'b', 'p', 'k', and single fricatives, as 'f', 'S', are affected more positively in terms of substitution error.

6. Conclusion

In this paper we experimented with a cross-corpus and cross-linguistic evaluation of a DNN-HMM phone recognizer which uses acoustic and reconstructed articulatory information. Results show that additional articulatory features always produce a phone error reduction w.r.t. the acoustic baseline (a DNN-HMM phone recognizer that only uses acoustic features) if the articulatory reconstruction is good enough. Furthermore the

analysis of error allowed us to identify the phonemic categories that mainly benefit from using reconstructed articulatory information across datasets and languages.

7. References

[1] Badino, L., Canevari, C., Fadiga, L. and Metta, G., "Deep-level acoustic-to-articulatory mapping for DBN-HMM based phone recognition", in Proceedings of IEEE SLT 2012, Miami, Florida,2012.

[2] Badino, L., Canevari, C., Fadiga, L. and Metta, G., Deep-Level Acoustic-to-Articulatory Mapping for DBN-HMM Based Phone Recognition - Erratum. Available at http://www.rbcs.iit.it/online/badino_et_al_slt2012_erratum.pdf

[3] Canevari, C., Badino, L., Fadiga, L., Metta, G., "Relevance-weighted reconstruction of articulatory features in Deep Neural Network-based Acoustic-to-Articulatory Mapping", in Proceedings of Interspeech, Lyon, France, 2013.

[4] Castellini C, Badino L, Metta G, Sandini G, Tavella M, Grimaldi M, Fadiga L., "The Use of Phonetic Motor Invariants Can Improve Automatic Phoneme Discrimination", in PLoS ONE 6(9): e24055. doi:10.1371/journal.pone.0024055.

[5] Cetin, O., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., and Livescu, K. (2007), An Articulatory Feature-Based Tandem Approach and Factored Observation Modeling, in IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, April 2007.

[6] Dahl, G. Yu, D., Deng, L. and Acero, A., "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30-42, January 2012

[7] Eide, E., Distinctive features for use in an automatic speech recognition system, in Proceedings of Eurospeech, Aalborg, Denmark, pp. 16131616. 2001.

[8] Fukuda, T., Yamamoto, W., and Nitta, T., , Distinctive phonetic feature extraction for robust speech recognition, in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 2528, 2003.

[9] Grimaldi, M., Gili Fivela, B., Sigona, F., Tavella, M., Fitzpatrick, P., Craighero, L., Fadiga, L., Sandini, G., Metta, G., "New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph", in Proceedings LangTech., Rome, Italy, 2008.

[10] Hinton, G.E., "Training products of experts by minimizing contrastive divergence", Neural Computational, vol. 14, pp. 1771-1800, 2002.

[11] Hinton, G.E., Osindero, S. and Teh, Y., "A fast learning algorithm for deep belief nets", Neural Computation, vol. 18, pp. 1527–1554, 2006.

[12] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M.,"Speech production knowledge in automatic speech recognition", J. of the Acoust. Soc. Am., vol. 121(2), pp. 723-742, 2007.

[13] Markov, K., Dang, J. and Nakamura, S., "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework" Speech Communication, 48, 161–175, 2006.

[14] Mohamed, A. R., Dahl, G. E. and Hinton, G. E."Deep belief networks for phone recognition", NIPS 22, workshop on deep learning for speech recognition, 2009.

[15] Mohamed, A., Hinton, G. E. and Penn, G. Understanding how Deep Belief Networks perform acoustic modeling. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. 4273-4276.2012.

[16] Ozbek, I.Y., Hasegawa-Johnson, M. and Demirekler, M., "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) with Audio-Visual Information Fusion and Dynamic Kalman Smoothing", IEEE Transactions on Audio, Speech, and Language 19(5):1180-1195, 2011.

[17] Richmond, K., King, S. and Taylor, P., Modeling the uncertainty in recovering articulation from acoustics, Computer Speech and Language, vol. 17(2), pp. 153-172, 2003.

[18] Richmond, K., ”Preliminary inversion Mapping Results with a New EMA Corpus”, in Proceedings of Interspeech, Brighton, UK, 2009.

[19] Richmond, K., Hoole, P., King, S., ”Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory corpus”, in Proceedings of Interspeech, Florence, Italy, 2011.

[20] Uria, B., Murray, I., Renals, S., Richmond, K., Deep architectures for articulatory inversion, In Proc. Interspeech, Portland, Oregon, USA, September 2012.

[21] Wrench, A.A. and Richmond, K., ”Continous speech recognition using articulatory data”, in Proceedings of the International Conference on Spoken Language Processing, pp. 145-148, 2000.

[22] Wrench, A. A., ”Multi-Channel/Multi-Speaker Articulatory Database for Continuous Speech Recognition Research”, Phonus., 5 . pp. 1-13, 2000.

[23] Available at http://data.cstr.ed.ac.uk/mocha/

[24] Available at http://htk.eng.cam.ac.uk/

[25] Available http://www.speech.cs.cmu.edu/SLM/toolkit.html