Relevance-Weighted-Reconstruction of Articulatory Features in Deep-Neural-Network-Based Acoustic-to-Articulatory Mapping

Claudia Canevari, Leonardo Badino, Luciano Fadiga, Giorgio Metta

RBCS, Istituto Italiano di Tecnologia, Genova, Italy

10.21437/Interspeech.2013-346

Abstract

We present a strategy for learning Deep-Neural-Network (DNN)-based Acoustic-to-Articulatory Mapping (AAM) functions where the contribution of an articulatory feature (AF) to the global reconstruction error is weighted by its relevance. We first empirically show that when an articulator is more crucial for the production of a given phone it is less variable, confirming previous findings. We then compute the relevance of an articulatory feature as a function of its frame-wise variance dependent on the acoustic evidence which is estimated through a Mixture Density Network (MDN). Finally we combine acoustic and recovered articulatory features in a hybrid DNN-HMM phone recognizer. Tested on the MOCHA-TIMIT corpus, articulatory features reconstructed by a standardly trained DNN lead to a 8.4% relative phone error reduction (w.r.t. a recognizer that only uses MFCCs), whereas when the articulatory features are reconstructed taking into account their relevance the relative phone error reduction increased to 10.9%.

Index Terms: Acoustic-to-Articulatory Mapping, critical articulators, Deep Neural Networks, phone recognition

1. Introduction

The trajectories, velocities and accelerations of the vocal tract articulators used as additional observations in a speech recognition system can significantly increase recognition accuracy ([1], [7]).

During recognition, when measured articulatory features are used as observations, an inverse procedure, called Acoustic-to-Articulatory Mapping (AAM), is required to recover the articulatory features (AFs) as only speech acoustics are available. One of the most used and best performing machine learning strategy to learn the AAM is the Multilayer Perceptron (MLP) [8]. Although an MLP cannot properly handle the non-uniqueness of the AAM problem, i.e., the fact that a given speech sound can be produced using a range of different vocal tract configurations [6], it can successfully capture its main nonlinearities.

Typically one single MLP is used to recover all the articulatory features given the acoustic spectrum. That means that the error function that has to be minimized during the MLP training is a function of the overall reconstruction error. Here we weight the contribution of each articulatory feature to the global reconstruction error function according to the relevance that each articulatory feature has in the production of a given speech sound. The motivation behind this relevance-weighted reconstruction is to improve the reconstruction of the articulators that are most critical for the production of a given phone to the detriment of the reconstruction of the articulators that are less critical.

In this study the relevance of an AF is computed as a function of the variance it shows when a given sound is produced. Indeed empirical evidence shows that the less critical an articulator the more variable its behavior ([10], [12]).

Here we experiment with two methods to evaluate the relevance of an AF in the production of a phone. The first method measures the variance of the actual AF in each phone state whereas the second method computes, through Mixture Density Networks [3], the estimated frame-wise variance given the acoustic evidence, one for each articulatory feature. The first method has the only aim to verify if our AF relevance evaluation is reasonable and effective since it considers directly the actual AFs. Note that in the second method, contrary to [11] where MDNs are directly used to carry out the AAM, the MDNs are used in an intermediary step to collect information that is then used to train the MLP that will perform AAM.

It is important to stress that the relevance-based weights applied to the backpropagation error function act only on the hidden layers, but not on the learning of the parameters of the output layer. In fact the output layer is a set of regressors (one for each AF) that are independent given the output of the last hidden layer (which encodes a shared representation of the input).

Similar to [1] and [13] the MLP we use to perform the AAM is actually a Deep Neural Network (DNN), in other words an MLP whose parameters are first pretrained by unsupervised pretraining of its corresponding Deep Belief Network [4]. Once the AFs are recovered, they are combined with acoustic features and used as observations in a hybrid DNN-HMM based phone recognizer as in [1]. DNN-HMM systems are the state-of-the-art in phone recognition [9], which makes our acoustic baseline (i.e., DNN-HMM system that only uses MFCCs) a very strong baseline.

2. Mixture Density Networks

A Mixture Density Network (MDN) combines a standard Multilayer Perceptron (MLP) with a mixture model [2]. The MLP takes the input vector x and maps it into a mixture model control parameters vector which generates the conditional probability density function of the target variable t of dimension L. Here the mixture model is a Gaussian Mixture Model (GMM).

$$p(t|x) = \sum_{k=1}^{K} \pi_k N(t|\mu_k(x), \sigma_k^2(x)) \quad (1)$$

where $\pi_k(x)$ are the mixing coefficients, $\mu_k(x)$ are the (L-dimensional) means, $\sigma_k^2(x)$ are the spherical variances and K is the number of components of the GMM.

The MLP has sigmoidal hidden units and one output unit for each control parameter. The output layer consists of K softmax

output units, K exponential output units and KxL linear output units which determine the mixing coefficients, the spherical variances and the components of the means respectively. The weights and biases of the MLP are adjusted to minimize the following error function through standard backpropagation.

$$e = -\sum_{i=1}^{I} ln(\sum_{k=1}^{K} \pi_k(x_i, w) N(t_i|\mu_k(x_i, w), \sigma_k^2(x_i, w))) \quad (2)$$

Once the GMM parameters are learnt we can compute the conditional expectation of t:

$$E[t|x] = \sum_{k=1}^{K} \pi_k(x)\mu_k(x) \quad (3)$$

and the overall variance of the conditional distribution $p(t|x)$ [3]:

$$s^2(x) = \sum_{k=1}^{K} \pi_k(x)[\sigma_k^2(x) + ||\mu_k(x) - \sum_{l=1}^{K} \pi_l(x)\mu_l(x)||^2] \quad (4)$$

In this work the estimated variance is used to compute the frame-wise AF relevance given the acoustic evidence. In addition the MDN represents an alternative AAM method in which an MDN is used to recover each AF (as in [11]). In both cases the MDNs have the same configuration (see section 4).

3. Deep Neural Networks

A Deep Neural Network (DNN) is an MLP whose parameters are pretrained by unsupervised training of a "corresponding" Deep Belief Network (DBN). The pretraining can be interpreted as an attempt to extract the statistical structure of the input domain (i.e., $P(X)$) that can effectively guide the search of input-output relations ($P(Y|X)$) in classification or regression problems.

In DNN training, a DBN is first trained in an unsupervised fashion. Then the DNN is created by transforming the stochastic nodes of the trained DBN into deterministic nodes and by adding an output layer on top of the network. The output unit activation function can be either e.g. a linear regressor or a softmax function depending on whether we want the DNN to perform regression or classification. Finally the parameters of the DNN are fine-tuned through supervised learning, typically using backpropagation.

In this work the DNNs are used (i) to learn the AAM and (ii) to estimate the phone posteriors given the acoustic and articulatory evidence.

In order to weight the importance of the reconstruction errors according to the relative relevance of each AF in the production of a given speech sound, the backpropagation cost function, i.e. the mean square reconstruction error, is multiplied by a weight matrix H:

$$E = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} ((y_{ij} - t_{ij})h_{ij})^2 \quad (5)$$

where y_{ij} is the recovered j-th AF in the i-th frame, t_{ij} is the actual AF, h_{ij} is the relevance of the j-th AF in the i-th frame (see section 5 for more details on AF relevance evaluation), I is the number of samples and J is the number of AFs.

4. Experimental setup

We used the 460 British-English utterances of the msak0 male voice of the MOCHA-TIMIT corpus [16]. They consist of simultaneous recordings of speech and Electromagnetic Articulographic (EMA) data (plus other articulatory data that we did not consider). Training and testing were carried out using the same 5-fold cross-validation as in [15] and [1]. Acoustic and articulatory features were extracted as described in [1]. We used vectors of 60 mel-filtered spectral coefficients (MFSCs) as acoustic input for the AAM, vectors of 39 MFCCs (first 12 MFCCs and energy coefficients, plus deltas and delta-deltas) as acoustic observations in the DNN-HMM phone recognition system and 42 AFs consisting of the x and y trajectories, plus their first and second derivatives, of upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue blade (TB) and tongue dorsum (TD). Note that the upper incisors exhibit very small variations in all phones and are used for head-movement correction. Nevertheless they can provide important relative information on how an articulator is positioned with respect to the others. Here upper incisors are reconstructed from speech acoustics as the other articulators, but they are not considered during the AF relevance evaluation.

The AAM was either performed by a DNN or a set of MDNs (whose outputs could be used to reconstruct the AFs or to compute the relevance of each AF). The DNN is a 3-hidden layer net with 300 nodes per each hidden layer. The input units of the corresponding DBN were Gaussian-distributed while all hidden units were binary. The MDNs were 1-hidden layer nets with 60 hidden units and 3 spherical mixture components. We trained one MDN for each AF. Both DNN and MDNs have an input of 5 acoustic features vectors (60 x 5 MFSCs) and the output is the vector of 42 AFs corresponding to the frame on which the acoustic input is centered.

We used 3 states per phone. The state boundaries were computed in the training utterances using the HInit, HRest and HERest functions of the HTK [17].

The phone posteriors were computed by a 3-hidden layer DNN, with 9 vectors of MFCCs (39 x 9 MFCCs) and the corresponding 9 vectors of AFs (42 x 9 AFs), when AFs were used, as input units. Each hidden layer has 1500 units while the output layer has 132 units (44 British English phonemes of MOCHA-TIMIT x 3 states). The estimated phone posteriors (not divided by the phone priors), plus phone unigrams and bigrams and state bigrams (all computed from the training data) were fed into a Viterbi decoder that computed the most probable phone sequence of each test utterance.

5. Articulatory feature relevance

We assumed that the relevance of an AF is a function of the variance it shows when a given sound is produced.

In order to qualitatively explore the relevance-variance relation, we plotted the x-y positions of the 7 vocal tract points during the production of different phones. Figures 1 and 2 show the x-y plots of the positions of the 7 vocal tract points for the */b/* and */dh/* phonemes. Comparing the two plots we can observe that the distributions of positions of the lips are smaller in the labial sounds (where the lips are the most critical articulators) than in the dental sounds. On the other hand, the distributions of the tongue positions are smaller in the dental sounds (where the tongue and particularly the tongue tip is the most critical articulator) than in the labial sounds.

In order to quantitatively evaluate the relevance of an AF

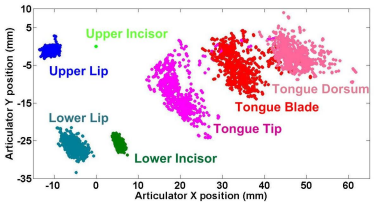

Figure 1: *Plots of the positions of the 7 vocal tract points in the midsagittal plane over all the /b/ phones in the MOCHA-TIMIT msak0 data.*

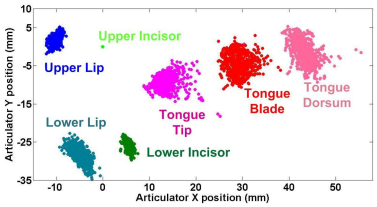

Figure 2: *Plots of the positions of the 7 vocal tract points in the midsagittal plane over all /dh/ phones in the MOCHA-TIMIT msak0 data.*

we experimented with two methods. In the first method, for each phone state we measured the variance of each actual AF. In the second method we estimated the frame-wise AF variance given the acoustic evidence computed by an MDN (equation 4).

The first method has the only aim to explore the suitability of the relevance-variance relation. In fact it cannot be used in practice as it needs to directly consider the actual AF values in their phone states.

In the second method each MDN estimates the frame-wise variance of a reconstructed AF without using any prior phonological knowledge.

From each of the two variances computed with the two methods we derived two different measurements of AF relevance. One is simply the inverse of the (actual state-level or MDN estimated frame-level) standard deviation. The other is the inverse of the ratio between the (actual state-level or MDN estimated frame-level) standard deviation and the overall standard deviation of the AF. The first is an absolute estimation of AF relevance whereas the second is a relative estimation.

We used four indexes to identify the types of AF relevance. **m0** and **m1** respectively refer to the absolute and the relative relevance computed from the actual state-level variance. **e0** and **e1** respectively refer to the absolute and the relative relevance computed from the estimated frame-level variance. All the relevance values were mapped into a [1 10] range.

Four different types of the weight matrix H of equation 5 were built using the four types of AF relevance.

6. Results

6.1. Articulatory reconstruction

The AF reconstruction is evaluated using the Root-mean-square error (RMSE) and the Pearson product moment correlation coefficient (r).

$$RMSE_f = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(o_{f,i} - t_{f,i})^2} \qquad (6)$$

$$r_f = \frac{\sum_{i=1}^{N}(o_{f,i} - \bar{o}_f)(t_{f,i} - \bar{t}_f)}{\sqrt{\sum_{i=1}^{N}(o_{f,i} - \bar{o}_f)\sum_{i=1}N(t_{f,i} - \bar{t}_f)}} \qquad (7)$$

where N is the number of frames in the testing set, $o_{f,i}$ is the estimated value for the AF f in the i-th frame, $t_{f,i}$ is the actual value of f in the i-th frame, $\bar{o}_f$ and $\bar{t}_f$ are the mean of the estimated and actual value of f respectively. Both RMSE and r are averaged over all the AFs.

A first comparison between the different AAM methods (Table1) shows that weighting the backpropagation error function slightly improves the overall AF reconstruction (both in terms of RMSE and r). The weights computed from the frame-level MDN-estimated-variance outperform the weights computed from the actual state-level variance and produce a 1.3 % significant relative RMSE reduction (w.r.t. the DNN baseline, where all AFs have the same importance).

Figures 3 and 4 show how the relative RMSE reduction is stronger for the trajectories of the critical AFs (lips for the '/b/' labial phoneme and the tongue for the '/dh/' dental phoneme). Note that a small RMSE reduction is observed for all articulatory features when the backpropagation error function is multiplied by a constant matrix H (this was verified using a constant value equal to 10). That explains why a very small reduction is shown even for some non-critical AFs in figures 3 and 4. This might be due to the fact that, for this particular dataset, when using a constant matrix H with constant >1, the reconstruction error slightly decreases.

Note that a reduction of the overall reconstruction error was not expected. In fact, the goal of the relevance weighted reconstruction is to improve the reconstruction of what matters (i.e., the critical AFs) to the detriment of the reconstruction of what is less important (the non-critical AFs).

The MDNs performed significantly worse than the DNN. The MDNs whose results are reported in Table 1 had 3 mixtures and were trained for 400 epochs. We tried a larger number of training epochs and different numbers of mixture but that resulted in a higher reconstruction error. That does not exclude that MDNs with a larger number of hidden layers and/or hidden units or even a pre-trained MDN as in [14], might perform better than a DNN. However the goal of the present study was not to compare MDN and DNN performances for AAM but to explore the utility (for acoustic-articulatory modeling) of a relevance weighted-reconstruction. The MDN was a tool to compute the variance-based relevance but also alternative techniques could be used, ideally techniques that require less training time. For example, even MDNs trained with many less training epochs might be equally successful if a gross estimation of the variance is sufficient (this is something we have not explored).

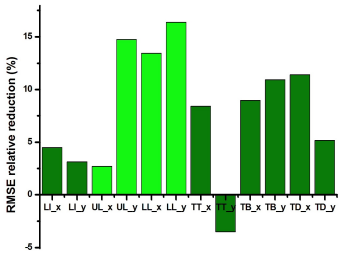

Figure 3: *RMSE relative reduction of each recovered articulatory trajectory using DNN_He0 w.r.t. the DNN baseline for phoneme '/b/'. The light green bars refer to the most critical AFs. The overall relative RMSE reduction is 4.68%.*

AAM	RMSE	r
DNN_baseline	0.693	0.692
MDNs	0.737	0.642
DNN_Hm0	0.690	0.692
DNN_Hm1	0.689	0.693
DNN_He0	**0.684**	0.697
DNN_He1	0.688	0.692

Table 1: *Articulatory reconstruction results averaged over the 5 splits for the following AAM methods: a) the DNN baseline in which all AFs have the same importance, b) the MDNs, one for each AF, c) the DNNs in which the weight matrix H is built computing* (i) *the inverse of the actual state-level standard deviations* (DNN_Hm0)*,* (ii) *the inverse of the ratio between the actual state-level standard deviations and the overall actual standard deviations* (DNN_Hm1)*,* (iii) *the inverse of the MDN estimated frame-level standard deviations* (DNN_He0) *and* (iv) *the inverse of the ratio between the MDN estimated frame-level standard deviations and the overall MDN estimated standard deviations* (DNN_He1)*. The value in bold represents a statistically significant reduction in RMSE compared to the DNN baseline* ($p < 0.05$*, one-tailed t-test*).

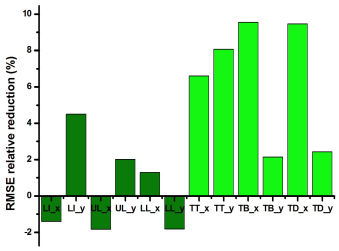

Figure 4: *RMSE relative reduction of each recovered articulatory trajectory using DNN_He0 w.r.t.the DNN baseline for phoneme '/dh/'. The light green bars refer to the most critical AFs. The overall relative RMSE reduction is 3.4%.*

6.2. Phone recognition

Table 2 shows the frame-wise classification accuracy and the phone error rate (PER) for the DNN-HMM phone recognition system using different types of observations. The recovered AFs always improve phone recognition. The PER reduction ranges from 8.4% to 10.9%. A perfect articulatory reconstruction would lead to a 26.4% PER reduction. The DNN-based relevance-weighted reconstruction produces up to an average 2.7% PER over a DNN-based reconstruction where all AFs are assumed to have the same relevance (DNN baseline). In the comparison between DNN_He0 and DNN_baseline we observed an error reduction in all cross-validation fold (figure 5).

Interestingly, the MDN recovered AFs produced a smaller PER than the DNN_baseline recovered AFs, although the overall MDN-based reconstruction was significantly worse than any DNN-based reconstruction. Such a mismatch raises the question on whether the overall RMSE and Pearson product moment correlation are the most appropriate metrics to evaluate the reconstruction of AFs for the ultimate goal of acoustic-articulatory modeling. Metrics (including modified versions of RMSE and r) that dynamically take into account the phonetic relevance of each AF seem to be more appropriate.

System	Features set	FwCa %	PER
GMM-HMM	MFCCs		38.0
DNN-HMM	MFCCs	65.9	32.2
DNN-HMM	MFCCs + actual AFs	73.9	23.7
DNN-HMM	MFCCs + DNN_baseline-RAFs	69.8	29.5
DNN-HMM	MFCCs + MDNs-RAFs	69.5	29.2
DNN-HMM	MFCCs + DNN_Hm0-RAFs	70.0	29.1
DNN-HMM	MFCCs + DNN_Hm1-RAFs	70.0	29.0
DNN-HMM	MFCCs + DNN_He0-RAFs	70.3	28.7
DNN-HMM	MFCCs + DNN_He1-RAFs	70.1	28.9

Table 2: *Frame-wise phone classification accuracy* $(FwCa)$ *and phone error rate* (PER) *using MFCCs only, MFCCs and actual AFs, MFCCs and recovered AFs* $(RAFs)$ *through the 6 methods previously described. RAFs are both in training and testing set when they are used. Values are averaged over the 5 splits. For comparison with previous work the first row shows the PER of the Gaussian Mixture Model-Hidden Markov Model baseline of [15] (where recovered AFs did not produce any improvement over the baseline). Note that we used the same DNN-HMM phone-recognizer as in [1]. However PER (and FwCa) are much smaller (larger) than in [1]. This is due to a longer pretraining of all DNNs and bug fixing of the Viterbi decoder.*

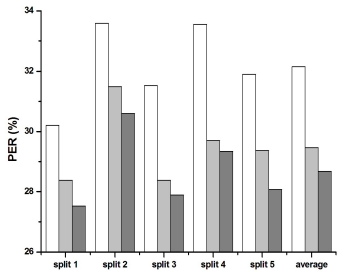

Figure 5: *Phone error rate (%) for each fold and on average using the following feature sets: MFCCs (white), MFCCs + DNN baseline - RAFs (light grey), MFCCs + DNN_He0-RAFs (dark grey)*

7. Conclusion

The vocal tract articulators are not all equally important to produce a given speech sound. When reconstructing the articulatory features by minimizing a global reconstruction error it may useful to try to force a more accurate reconstruction of the critical articulatory features to the detriment of the less important reconstruction of the articulatory features. In this paper we experimented with relevance-weighted reconstruction error functions for the supervised learning of Deep Neural Networks that perform Acoustic-to-Articulatory Mapping. The relevance of the articulatory features was computed as a function of the variance that the articulatory features show when producing a given sound. A frame-level variance was computed using Mixture Density Networks. Results show that our relevance-weighted reconstruction significantly reduced the overall reconstruction error and reduced the phone error rate of a hybrid Deep Neural Network/Hidden Markov Model phone recognizer that uses both acoustic and (reconstructed) articulatory features.

8. References

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

[2] Bishop, C. M., "Mixture density networks", Technical Report NCRG/4288, Neural Computing research Group, Department of Computer Science, Aston University, Birmingham, B4 7 ET, UK, February, 1994.

[3] Bishop, C. M., "Pattern recognition and machine learning", Springer Science+Business Media, LLC, 233 Spring street, New York, NY 10012, USA, 2009.

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

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

[6] Lindblom, B., Lubker, J. and Gay, T., "Formant frequencies of some fixed-mandible vowels and model of speech motor programming by predictive simulation", Journal of Phonetics, vol. 7, pp. 146-161, 1979.

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

[8] Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E. and Goldstein, L., "Retrieving tract variables from acoustics: a comparison of different machine learning strategy", IEEE J. of Selected Topics in Signal Processing, vol. 4(6), pp. 1027-1045, 2010.

[9] Mohamed, R., Dahl, G. E. and Hinton, G. E., "Deep belief networks for phone recognition", NIPS 22, work-shop on deep learning for speech recognition, 2011.

[10] Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zachs, J. and Levy, S., "Inferring Articulation and Recognising Gestures from Acoustics with a Neural Network Trained on X-ray Microbeam Data", The Journal of the Acoustical Society of America, 92(2), pp. 688-700, 1992.

[11] 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.

[12] Rose, R. C., Schroeter, J. and Sondhi, M. M., "The potential role of speech production models in automatic speech recognition", The Journal of the Acoustical Society of America, 99(3), pp. 1699-1709, 1996.

[13] Uria, B., Renals, S. and Richmond, K., "A deep neural network for acoustic-articulatory speech inversion", In Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning, Sierra Nevada, Spain, 2011.

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

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

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

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