

# IMPROVING GENERALIZATION OF VOCAL TRACT FEATURE RECONSTRUCTION: FROM AUGMENTED ACOUSTIC INVERSION TO ARTICULATORY FEATURE RECONSTRUCTION WITHOUT ARTICULATORY DATA

*Rosanna Turrise, Raffaele Tavarone, Leonardo Badino*

CTNSC, Istituto Italiano di Tecnologia, Ferrara, Italy

## ABSTRACT

We address the problem of reconstructing articulatory movements, given audio and/or phonetic labels. The scarce availability of multi-speaker articulatory data makes it difficult to learn a reconstruction that generalizes to new speakers and across datasets. We first consider the XRMB dataset where audio, articulatory measurements and phonetic transcriptions are available. We show that phonetic labels, used as input to deep recurrent neural networks that reconstruct articulatory features, are in general more helpful than acoustic features in both matched and mismatched training-testing conditions. In a second experiment, we test a novel approach that attempts to build articulatory features from prior articulatory information extracted from phonetic labels. Such approach recovers vocal tract movements directly from an acoustic-only dataset without using any articulatory measurement. Results show that articulatory features generated by this approach can correlate up to 0.59 Pearson’s product-moment correlation with measured articulatory features.

**Index Terms**— Articulatory features, tract variables, acoustic inversion, deep learning, XRMB

## 1. INTRODUCTION

Measurements of vocal tract movements can be beneficial for several speech technology applications, including speech synthesis [1], automatic speech recognition (ASR) [2, 3], pronunciation training [4] and speech-driven computer animation

[5]. Typically, vocal tract movements, henceforth referred to as articulatory features (AFs), are much more difficult to collect than audio and require extensive preprocessing steps to reduce noise and interpolate missing data [6]. This results in few and relatively small corpora of articulatory data and, as a consequence, in a strong limitation to their use in most of the aforementioned cases. Learning a reliable AF reconstruction, that generalizes well across speakers and datasets, would allow a more significant use of articulatory information in many applications. Previous works on AF reconstruction learn an acoustic inversion (AI), i.e., a mapping from acoustic features to AFs (e.g., [7, 8, 9]). While most of these studies have focused on speaker-dependent AI, there is some recent work on the speaker-independent case [10, 11, 12].

In this paper we address two questions: (1) is that possible to learn an AI that better generalizes to new speakers by either augmenting or substituting altogether the acoustic input with some phonetic information? (2) Can we generate accurate AFs, starting from some phone-specific prior articulatory knowledge and using very little or zero vocal tract measurements?

To address question 1 we use input phonetic features that range from phone labels to phonological features, which can be extracted from those labels through a look-up table. Specifically, we use phonological features from the Articulatory Phonology theory [13, 14]. Although the idea of pairing phone labels with input acoustic features to recover AFs is not new [15, 16], here we test

the utility of phonological features in both matched and mismatched training-testing conditions. The mismatched condition is created within the XRMB dataset [17] by training and validating on male speakers and testing on female speakers, and vice versa. We expect the phonetic information to be particularly helpful in the mismatched case, as it is speaker and environment independent.

Henceforth we will refer to AI and its variants as supervised methods, in which measured articulatory data are used as targets to train a bidirectional long short-term memory recurrent network (LSTM) to perform AF reconstruction.

Adding side information, as proposed here, or using adaptation techniques to make AF reconstruction more general may still be very challenging as existing articulatory datasets are small and only cover the read-speech speaking style.

A possible alternative, explored in this paper, is to extract AFs directly from audio-only datasets given weak prior knowledge about average vocal tract configurations typical of each phoneme. This alternative strategy addresses our question 2 and the proposed methods are defined as weakly supervised. This approach in principle does not require any articulatory data but some articulatory data can still be used to compute or refine the articulatory priors (hence the name “weakly supervised”).

Our 3 weakly supervised methods are based on deep auto-encoders [18, 19] or residual networks [20] and tested on the XRMB dataset. Phone-dependent discrete articulatory priors, extracted from phonemes through a look-up table, are used to generate real-valued latent articulatory representation of the acoustic data.

## 2. ARTICULATORY FEATURES

We considered the following AF sets:

**Pellet trajectories (PTs).** Preprocessed x-y trajectories of 8 pellets tracking speaker’s lips, tongue and jaw (see [11] for preprocessing details).

**Vocal Tract Variables (VTVs),** from articulatory phonology theory [13, 14]. Specifically, we considered lip protrusion (LP) and aperture (LA), tongue tip constriction location (TTCL) and degree

(TTCD), tongue body constriction location (TBCL) and degree (TBCD). The 6 VTVs were extracted from pellet trajectories by using the transformation procedure described in [21]. The extraction requires the shape estimation of the hard palate, which was computed by fitting a second-degree polynomial curve to the tongue measurement data.

**Phone-dependent extended discrete VTVs: LFs and SFs.** To each phoneme we assigned one vector consisting of 10 integer-valued features: the aforementioned 6 VTVs, 2 additional manually annotated vocal tract features (specifically, velic opening degree (VEL) and glottal opening degree (GLO)), consonant, and silence. We used two different feature sets: LFs and SFs. LFs refers to the set where the values (integers) of the first 6 VTVs were provided by an expert, while in SFs set they were computed through a simple statistical procedure. For each phone label we computed the average values of the per-speaker z-normalized VTVs over the XRMB training dataset. Average values were then rounded to their closest integer, resulting in an average number of 4 quantization levels per feature (while LFs have on average 5 levels per feature). Both LFs and SFs can be retrieved from each phoneme through a look-up table (available *here.*), so we refer to them as phonological features.

## 3. SUPERVISED METHODS

Supervised methods rely on datasets consisting of audio, phonetic annotations and measured articulatory data. The goal is to learn a mapping from acoustic features (e.g., mel-scaled frequency cepstral coefficients, MFCCs) and/or phonological features (i.e., phone labels or LFs or SFs) to AFs (either in the form of PTs or VTVs). In our experiments these mappings are learned by training deep bidirectional recurrent neural network based on Long short-term memory (LSTM) cells [22].

## 4. WEAKLY SUPERVISED METHODS

In this strategy the available articulatory information consists of some prior concise description of the typical vocal tract configuration of each phone

(independent of the phonetic context). These priors are either provided by an expert (LFs), or empirically extracted from some training articulatory data (SFs). We experimented with SFs extracted from multiple-speaker data and single speaker-data. In this section, we denote by  $\mathbf{x}$  the acoustic feature vector, by  $\hat{\mathbf{x}}$  the reconstructed acoustic feature vector, by  $\mathbf{z}$  the articulatory prior vector (i.e., SFs or LFs) and by  $\hat{\mathbf{z}}$  the generated AF vectors. The precision of the generated articulatory features is evaluated by comparing  $\hat{\mathbf{z}}$  with measured articulatory features.

#### 4.1. Autoencoder-based method

An autoencoder (AE) is an artificial neural network architecture that attempts to reconstruct its input through a latent representation (encoding). It consists of two parts: a mapping from the input to the latent representation (encoder,  $e$ ), and the input reconstruction starting from the encoding (decoder,  $d$ ).

##### 4.1.1. Autoencoder 1

The first autoencoder (AE1) we propose takes the audio as input and returns its reconstruction. This map goes through the encoding layer, which we would like to resemble an articulatory representation by adding an additional term to the standard autoencoder. Let  $\mathbf{z}_t$  be the prior vector at time  $t$  with dimensionality  $G$  ( $G = 10$ , as mentioned above), and  $\mathbf{x}_{t-T}^{t+T} = [\mathbf{x}_{t-T}, \dots, \mathbf{x}_t, \dots, \mathbf{x}_{t+T}]$  the input concatenation of the audio vectors, where  $T$  is the context window length on each side. The objective function at time  $t$  is:

$$L_{A1,t} = \|\mathbf{x}_{t-T}^{t+T} - \hat{\mathbf{x}}_{t-T}^{t+T}\|_2^2 + \lambda_z \cdot \|\mathbf{z}_t - \hat{\mathbf{z}}_t\|_2^2, \quad (1)$$

where  $\hat{\mathbf{z}}_t = e(\mathbf{x}_{t-T}^{t+T})$ ,  $\hat{\mathbf{x}}_t = d \circ e(\mathbf{x}_{t-T}^{t+T})$  and  $\lambda_z$  is a scalar hyperparameter that weights the importance of the second term of the loss. In other words, we force the latent representation of the acoustic features  $\mathbf{x}$  to be close to the typical configuration taken by vocal tract when the phoneme associated to  $\mathbf{x}$  is produced. The  $\mathbf{z}$  can be seen as the mean of a prior multivariate Gaussian distribution, while we

do not make any prior assumption regarding its covariance (contrary to variational autoencoders [23]). The assumption that actual AFs are roughly normally distributed around  $\mathbf{z}$  is also shared by the next approaches, and is supported by qualitative analysis we have carried out per each phone.

##### 4.1.2. Autoencoder 2

In the second variant, autoencoder 2 (AE2), we revert the AE structure previously described in Sec. 4.1.1. Now,  $\mathbf{z}$  is the input of the AE which provides the articulatory reconstruction  $\hat{\mathbf{z}}$ . We force the encoding layer to match the acoustic latent representation  $\mathbf{x}$ . Therefore, the loss function to minimize at time  $t$  is:

$$L_{A2,t} = \|\mathbf{z}_{t-T}^{t+T} - \hat{\mathbf{z}}_{t-T}^{t+T}\|_2^2 + \lambda_x \cdot \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2, \quad (2)$$

where  $\hat{\mathbf{x}}_t = e(\mathbf{z}_{t-T}^{t+T})$ ,  $\hat{\mathbf{z}}_t = d \circ e(\mathbf{z}_{t-T}^{t+T})$  and  $\lambda_x \in \mathbb{R}$  is an hyperparameter. Note that here the articulatory reconstruction  $\hat{\mathbf{z}}$  is not a function of the acoustic features as in AE1, but a direct function of the phonological features.

#### 4.2. Residual-based method

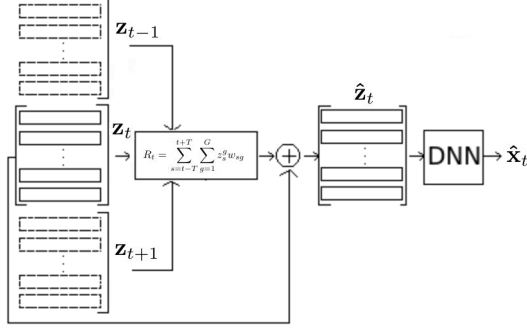
In this approach a deep neural network with one residual layer (ResDNN) takes articulatory prior vectors  $\mathbf{z}$  as input features and targets acoustic features (Figure 1). The residual layer [20] modulates the input  $\mathbf{z}$  with its left and right context weighted by a learned parameter, thus returning a coarticulation-modulated version of the  $\mathbf{z}$ .

Formally, the output of each  $i$ -th element of the residual layer  $\hat{\mathbf{z}}_t$  is defined as:

$$\hat{z}_t^i = z_t^i + R_t, \quad R_t = \sum_{s=t-T}^{t+T} \sum_{g=1}^G z_s^g w_{sg}^R. \quad (3)$$

$R_t$  is the residual at time  $t$ , and the  $w_{sg}^R$ 's are the learning parameters of the residual network. The sums taken over time and features model coarticulation effects. The network is trained to minimize the following loss function:

$$L_{R,t} = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2 + \lambda_w \|\mathbf{w}^R\|_2^2, \quad (4)$$



**Fig. 1:** Residual DNN structure. The frame context is only used in the residual layer. Here  $T = 1$ .

where  $\hat{x}$  is the reconstructed audio,  $\lambda_w$  controls the penalization term, and  $\mathbf{w}^R \in \mathbb{R}^{G \cdot (2T+1)}$ .

## 5. EXPERIMENTAL SETUP

### 5.1. Dataset

All the experiments were carried out on the 47 American English speaker subset of XRMB used in [11, 12], with the only exception that we discarded speaker *JW33* (used for validation in [11, 12]), as we discovered some corrupted audio (while we kept speaker *JW58* which was removed in [12] and only removed some corrupted utterances).

Articulatory data consists of x-y trajectories of: upper and lower lips, 4 tongue points, one mandible molar and one mandible incisor. For the training-testing matched condition we split the dataset into disjoint sets of 35/7/4 speakers for training/validation/testing respectively.

For the training-testing mismatched condition we split the dataset by gender. We refer to the so-obtained subsets as *Male* and *Female*, with 22 and 24 speakers respectively. For supervised methods, when *Female* was used as testing dataset, *Male* was split into 18/4 speakers for training/validation respectively. In the opposite case, *Female* was split into training and validation, with 19 and 5 speakers respectively.

Articulatory features were preprocessed as in [11], while acoustic features are the first 13

MFCCs, computed every 10ms from 25ms Hamming windows, plus deltas and delta-deltas. Both acoustic and articulatory features are per-speaker z-normalized.

### 5.2. Neural Networks

Supervised methods are based on bidirectional LSTMs (BLSTMs). The networks have 5 layers each containing 250 memory blocks, with peephole connections and hyperbolic tangent activation function. All experiments were carried out using Adaptive Momentum Optimizer [24], a piecewise constant learning rate with initial value set to 0.1, a 0.9 momentum,  $\epsilon = e^{-8}$  and initial decay rates of first and second moments 0.9 and 0.999, respectively. Weights were initialized with Xavier initialization [25]. Early stopping was applied to determine the number of training epochs.

In all weakly supervised methods, the network input consists of the central vector plus  $T = 12$  context vectors per side. Training was performed with stochastic gradient descent. Learning rate exponentially decayed every 10000 steps, with initial value 0.01 and 0.96 decay rate. Training was performed for 50 epochs or stopped earlier if the acoustic feature reconstruction error did not decrease.

Both AE types have a hourglass shape, symmetric w.r.t. the encoding layer. Each encoder (as well as the decoder) has 3 layers with 200, 130, 70 nodes respectively, decreasing towards the encoding layer which has  $G = 10$  nodes in AE1 and 39 nodes in AE2. Again we used Xavier initialization.

ResDNNs have 4 layers with 1000 nodes each, while the residual layer has  $G = 10$  nodes. We fixed  $\lambda_w = 0.01$  and grid-searched the remaining hyperparameters, based on the audio reconstruction.

We evaluated all methods by computing the average (over features) root mean squared error (RMSE) and the average Pearson's correlation coefficient ( $r$ ) between per-speaker z-normalized reconstructed and measured AFs (so RMSE is a normalized RMSE).

Input	PTs		VTVs	
	RMSE	$r$	RMSE	$r$
MFCCs (S1)	0.894	0.448	0.879	0.517
MFCCs	0.685	0.721	0.646	0.777
Phonemes	0.664	0.742	0.617	0.782
LFs	0.672	0.732	0.611	0.797
SFs	0.667	0.744	0.618	0.783
MFCCs + Phonemes	0.654	0.757	0.606	0.797
MFCCs + LFs	0.657	0.748	0.602	0.801
MFCCs + SFs	0.655	0.752	0.606	0.798

**Table 1:** Supervised methods results on the test set for PT and VTV reconstruction in the matched condition case. MFCCs (S1) refers to a BLSTM trained on 1 single speaker data (JW14).

## 6. RESULTS

### 6.1. Matched conditions

In Table 1, we compare the average RMSE and correlation for PT and VTV reconstruction of different BLTSM inputs. BLTSM training and evaluation were repeated twice, with different random initialization. To keep tables more readable we only report the mean, the std. dev. is always lesser than 0.01.

Results suggest that phonological features (phone labels, LFs and SFs) can outperform MFCCs, and, surprisingly, MFCCs slightly improve reconstruction when combined with phonological features, despite MFCCs containing much more detailed information than phonological features. LFs and SFs do not produce relevant improvement w.r.t. phone labels. Table 1 also shows AI results when only one speaker is used for training in order to quantify the gap w.r.t. to multi-speaker training data and to compare with weakly supervised methods in a limited articulatory data setting.

Results of weakly supervised methods in the matching conditions setting are summarized in Table 2. We compare them with the Baseline case, where the phonological features are directly compared with measured AFs. Again, all experiments were carried out twice (std. dev. < 0.01). Although LFs and SFs have a similar number of quantization levels, SFs largely outperform LFs in all methods.

Features	Baseline		ResDNN		AE1		AE2	
	RMSE	$r$	RMSE	$r$	RMSE	$r$	RMSE	$r$
LFs	-	0.366	-	0.360	-	0.330	-	0.390
SFs	0.858	0.524	1.010	0.554	0.862	0.507	0.820	0.571
SF1s	0.888	0.514	1.117	0.537	0.876	0.508	0.835	0.563
SF2s	0.872	0.519	1.102	0.524	0.894	0.492	0.826	0.568

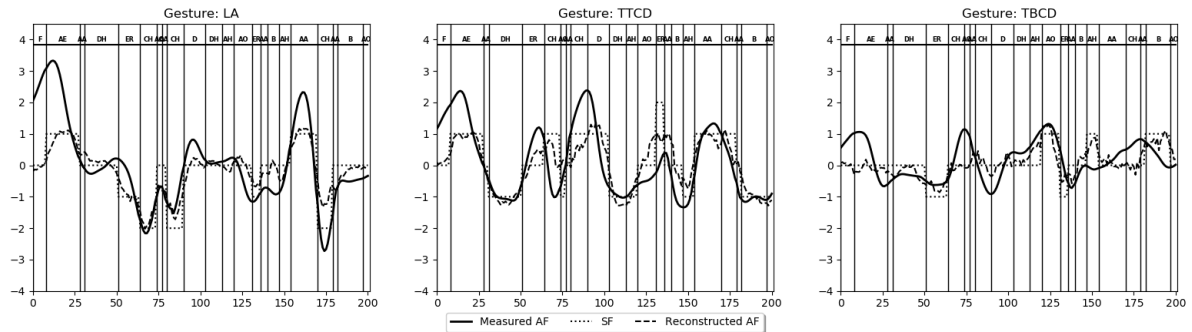
**Table 2:** Weakly supervised methods results on the test set. SF1s and SF2s refer to the statistical features computed on the JW14 and JW14+JW12 articulatory data, respectively.

		LP	LA	TTCL	TTCD	TBCL	TBCD
		RMSE	$r$	RMSE	$r$	RMSE	$r$
JW48	RMSE	0.825	0.859	0.838	0.753	0.816	0.828
	$r$	0.600	0.519	0.590	0.680	0.581	0.563
JW53	RMSE	0.781	0.842	0.845	0.666	0.739	0.745
	$r$	0.688	0.548	0.581	0.747	0.681	0.686

**Table 3:** Details of AE2 performance for speakers JW48 and JW53 (matched conditions).

Most importantly, the generated AFs  $\hat{z}$  always correlate more with actual AFs than the priors  $z$ , with the exception of method AE1. That means that AE2 and ResDNN successfully transform the original prior articulatory information into articulatory features that are closer to the actual AFs. AE2 is the most effective method.

To show that SFs well generalize across speakers, we re-computed the SFs based on only one or two training speakers (SF1s and SF2s) and repeated the weakly supervised experiments. Interestingly, results obtained with SF1s and SF2s do not significantly differ from SFs. This implies that the statistical representations calculated on few speakers (or just one!) well characterize the vocal tract of any other speaker. Importantly, in this limited data setting, ResDNN and AE2 outperform the best supervised method (e.g.,  $r = 0.537$  and  $r = 0.563$  vs.  $r = 0.517$ ). Note that Table 2 shows the best AE1 and AE2 performances on the validation set, achieved by fixing  $\lambda_z$  and  $\lambda_x$  at 2 and 0.5, respectively. We did not report the RMSE for the LFs, as they do not reflect the real measurements of the articulatory data. More detailed results can be found in Table 3, where the best AE2 performance is reported for two test speakers and for each VTV.



**Fig. 2:** Comparison between measured, statistical, and reconstructed (from AE2) LA, TTCD, TBCD features of speaker JW53.

Input	Test gender	RMSE	$r$
MFCCs	<i>Male</i>	0.848	0.592
SFs	<i>Male</i>	0.604	0.782
MFCCs + SFs	<i>Male</i>	0.685	0.743
MFCCs	<i>Female</i>	0.860	0.557
SFs	<i>Female</i>	0.625	0.787
MFCCs + SFs	<i>Female</i>	0.686	0.748

**Table 4:** BLSTM cross-gender VTV reconstruction.

Test gender	Baseline		AE2	
	RMSE	$r$	RMSE	$r$
<i>Male</i>	0.854	0.539	0.816	0.586
<i>Male (S1)</i>	0.877	0.526	0.822	0.579
<i>Female</i>	0.858	0.529	0.821	0.576
<i>Female (S1)</i>	0.867	0.529	0.819	0.576

**Table 5:** Cross-gender evaluation of AE2. *Male (S1)* and *Female (S1)* refer SFs computed from female speaker JW14 and male speaker JW12, respectively.

## 6.2. Mismatched conditions

Table 4 shows the results of the supervised methods in the training-testing mismatched conditions. The most striking result is that MFCCs not only perform significantly worse than SFs but even worsen SFs performance when combined with them. This is due to the strong

speaker dependency of MFCCs (despite their per-speaker normalization), that may be alleviated through speaker adaptation.

Regarding weakly supervised methods, in this case AE2 only, we computed SFs on one set and used them as prior information for training AE2 on the other dataset. Note that, in this case, AE2 is trained and tested on the same speakers (e.g., *Female*), while priors are computed on other speakers (e.g., *Male*). Indeed, since AE2 is only trained on acoustic data, which are always available, there is no need to generalize to new speakers. Results in Table 5 show that (i) AE2 almost matches the supervised method with MFCC; (ii) even in the mismatched case, AE2 reconstruction is not affected by a reduction of articulatory data to a single speaker.

## 7. CONCLUSIONS

In this paper we addressed articulatory feature reconstruction. We first showed that phone labels are more helpful than acoustic features in reconstructing AFs in both matched and mismatched conditions. We then proposed weakly supervised methods to reconstruct AFs from discrete articulatory priors extracted from phone labels. Results show that weakly supervised methods can be a more viable strategy when the amount of articulatory data is limited, especially in mismatched conditions.

## 8. ACKNOWLEDGEMENT

We thank Karen Livescu for providing the preprocessing XRMB data and the linguistic features. This work was partly supported by the EU’s Horizon2020 project ECOMODE (grant agreement No 644096).

## 9. REFERENCES

- [1] Richmond K. Yamagishi J. Ling, Z. H., “Articulatory control of hmm-based parametric speech synthesis using featurespace-switched multiple regression,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 207219, 2013.
- [2] Arora R. Livescu K. Bilmes J. Wang, W., “Unsupervised learning of acoustic features via deep canonical correlation analysis,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4590–4594, 2015.
- [3] Canevari C. Fadiga L. Metta G. Badino, L., “Integrating articulatory data in deep neural network-based acoustic modeling,” *Computer Speech and Language*, vol. 36, pp. 173195, 2016.
- [4] Bailly G. Badin P. Elisei F. Hueber, T., “Speaker adaptation of an acoustic-to-articulatory inversion model using cascaded gaussian mixture regressions,” in *14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, Aug. 2013, pp. 2753–2757.
- [5] Shimodaira H. Braude D. A. Ben-Youssef, A., “Articulatory features for speech-driven head motion synthesis,” Lyon, France, 2013.
- [6] Weiran Wang, Raman Arora, and Karen Livescu, “Reconstruction of articulatory measurements with smoothed low-rank matrix completion,” in *IEEE SLT*, Lake Tahoe, Nevada, USA, 2014.
- [7] K. Richmond, S. King, and P. Taylor, “Modelling the uncertainty in recovering articulation from acoustics,” *Computer Speech and Language*, vol. 17, no. 2, pp. 153172, 2003.
- [8] B. Uria, Murray, S. I., Renals, and K. Richmond, “Deep architectures for articulatory inversion,” in *Proc. of Interspeech*, Portland, Oregon, USA, 2012.
- [9] C. Canevari, L. Badino, L. Fadiga, and G. Metta, “Cross-corpus and cross-linguistic evaluation of a speaker-dependent dnn-hmm asr system using ema data,” in *Workshop on Speech Production for Automatic Speech Recognition*, Lyon, France, 2013.
- [10] Narayanan S. S. Ghosh, P. K., “An subject-independent acoustic-to-articulatory inversion,” Prague, Czech Republic, 2011.
- [11] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, “Unsupervised learning of acoustic features via deep canonical correlation analysis,” in *Proc. of ICASSP*, Brisbane, Australia, 2015.
- [12] Franceschi L. Arora R. Donini M. Pontil M. Badino, L., “A speaker adaptive DNN training approach for speaker-independent acoustic inversion,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, August 20-24, 2017, 2017, pp. 984–988.
- [13] Goldstein L. M. Browman, C. P., *Towards an articulatory phonology.*, 1986.
- [14] Goldstein L. Browman, C. P., “Articulatory phonology: An overview,” *Phonetica*, pp. 155–180, 1992.
- [15] Liu X. Wang L. Xie, X., “Deep neural network based acoustic-to-articulatory inversion using phone sequence information,” in *Interspeech 2016*, 2016, pp. 1497–1501.
- [16] Thomas Hueber, Atef Ben-Youssef, Gérard Bailly, Pierre Badin, and Frederic Elisei, “Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory hmm for pronunciation training,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [17] J.R. Westbury, *X-ray Microbeam Speech Production Database User’s Handbook: Version 1.0 (June 1994)*, Waisman Center on Mental Retardation & Human Development, 1994.
- [18] Huang J. C. Yang W. C. Liou, C. Y., “Modeling word perception using the elman network,” *Neurocomputing*, vol. 71, no. 16, pp. 3150 – 3157, 2008, Advances in Neural Information Processing (ICONIP 2006) / Brazilian Symposium on Neural Networks (SBRN 2006).
- [19] Cheng W. C. Liou J. W. Liou D. R. Liou, C. Y., “Autoencoder for words,” *Neurocomputing*, vol. 139, pp. 84 – 96, 2014.
- [20] Zhang X. Ren S. Sun J. He, K., “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.
- [21] Mitra V. Tiede M. Hasegawa-Johnson M. Espy-Wilson C. Saltzman E. Goldstein L. Nam, H., “A procedure for estimating gestural scores from speech acoustics,” *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3980–3989, 2012.
- [22] Schmidhuber J. Hochreiter, S., “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

- [23] Welling M. Kingma, D. P., “Auto-encoding variational bayes,” 12 2013.
- [24] Ba J. L. Kingma, D. P., “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
- [25] Bengio Y. Glorot, X., “Understanding the difficulty of training deep feedforward neural networks,” in *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, May 2010, vol. 9, pp. 249–256.