

# Learning robust speech representation with an articulatory-regularized variational autoencoder

Marc-Antoine Georges<sup>1,2</sup>, Laurent Girin<sup>1</sup>, Jean-Luc Schwartz<sup>1</sup>, Thomas Hueber<sup>1</sup>

<sup>1</sup>Univ. Grenoble Alpes, CNRS, GIPSA-lab, 38000 Grenoble, France

<sup>2</sup>Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France

marc-antoine.georges@grenoble-inp.fr

## Abstract

It is increasingly considered that human speech perception and production both rely on articulatory representations. In this paper, we investigate whether this type of representation could improve the performances of a deep generative model (here a variational autoencoder) trained to encode and decode acoustic speech features. First we develop an articulatory model able to associate articulatory parameters describing the jaw, tongue, lips and velum configurations with vocal tract shapes and spectral features. Then we incorporate these articulatory parameters into a variational autoencoder applied on spectral features by using a regularization technique that constrains part of the latent space to represent articulatory trajectories. We show that this articulatory constraint improves model training by decreasing time to convergence and reconstruction loss at convergence, and yields better performance in a speech denoising task.

**Index Terms:** Speech production, representation learning, variational autoencoder, articulatory model, speech enhancement.

## 1. Introduction

Motor and perceptuo-motor theories of speech perception involve internal motor simulation processes [1, 2] which may be particularly recruited when perceiving speech in adverse (e.g. noisy) conditions [3]. Similarly, most computational models of speech motor control rely on an explicit access to motor representations, first when recovering the motor commands required to reach an acoustic target (inverse internal model), and then when simulating the acoustic consequences of articulatory gestures (direct internal model) [4, 5]. Inspired by human cognition and neurophysiology, the integration of motor/articulatory priors and constraints in automatic speech processing systems has motivated several studies, for increasing the robustness of automatic speech recognition (ASR) systems in noise [6, 7, 8, 9], for allowing a better control of text-to-speech (TTS) synthesis [10], or for designing voice restoration and pronunciation training systems [11]. Modeling the complex relationships between phonetic targets, articulatory movements and speech acoustics is also essential to build a computational model of speech perception and production [12, 13].

With both these technological and fundamental research goals in mind, we deal in this paper with the automatic learning of latent representation from the raw speech audio signal. We focus on deep generative models and in particular on the variational autoencoder (VAE) model [14, 15] which can be seen as a probabilistic version of a deep autoencoder. The VAE was shown to be able to learn relevant latent representations by disentangling dimensions like speaker identity or phonetic features [16, 17]. This model has already been successfully used in a variety of speech processing applications, e.g. [18, 19, 20, 21]. In line with speech perception theory, we propose in this paper

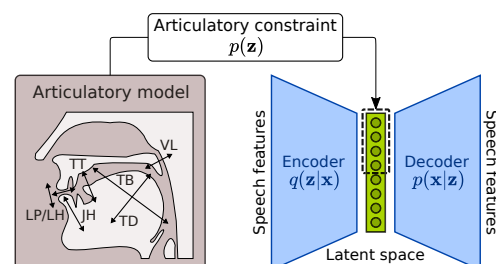


Figure 1: Schematic view of the proposed articulatory-regularized variational autoencoder.

to introduce prior articulatory information in the representation learning process. This information is first derived from an articulatory model built from in-vivo recordings of a reference speaker. It is then transferred at training time via an additional regularization term in the loss function of the VAE. An overview of the proposed architecture is shown in Figure 1. With this model, we address the two following research questions: 1) Can prior articulatory knowledge speed up the speech representation learning process? 2) Can prior articulatory knowledge make the learned latent representation more robust to noise? To address those questions, we compared the proposed articulatory-regularized VAE to a conventional one in terms of convergence speed at training time and on a speech denoising task.

To the best of our knowledge, the introduction of prior articulatory constraints for representation learning has been proposed only in very few studies. In [22], a set of vocal tract variables derived from the articulatory phonology theory [23] is used to constrain the latent space of a (deterministic) autoencoder. The performance is evaluated by measuring the accuracy of the reconstructed articulatory features whereas in the present study we focus on the quality of the reconstructed speech signal. In [24], a normalizing flow technique is used to constrain the latent space of two autoencoders respectively processing articulatory and audio data. However, evaluation is mostly limited to qualitative evaluation. Therefore, the present study proposes the first VAE model regularized by articulatory prior knowledge and used to learn robust latent representation from the audio speech signal.

## 2. Methodology

### 2.1. Acoustic and articulatory data

The following experiments were conducted on two datasets. Each of these datasets is composed of parallel audio and electromagnetic articulography (EMA) recordings (sustained vowels, vowel-consonant-vowel sequences, french words and sen-

tences). The first dataset, PB2007, consists of 1,109 items (15 minutes of speech) produced by a reference speaker (PB, male). EMA data were recorded using the Cartens 2D EMA system (AG200). Six coils were placed on the tongue tip, blade and dorsum, upper and lower lip and on the jaw (lower incisor). The acquired trajectories were low-pass filtered at 20 Hz and down-sampled from 200 Hz to 100 Hz. We denote by  $\mathbf{y}$  an individual resulting vector of EMA data, of dimension 12. The second corpus, BY2014,<sup>1</sup> includes 925 items (45 minutes of speech) produced by another reference speaker (BY, male). Articulatory trajectories were recorded using the 3D NDI Wave system with 9 coils used (3 on the tongue, 4 on the lips, 1 on the jaw, 1 on the velum). They were low-pass filtered at 20 Hz and down-sampled from 200 Hz to 100 Hz. The 3D coordinates of the 9 EMA coils were finally projected on the midsagittal plane resulting in a 14-dimensional vector  $\mathbf{y}$ .

As required by the LPCNet vocoder [25] which is used in this study to reconstruct an audio speech signal from the output of the VAE, each audio recording was converted into a sequence of 18 Bark-scale cepstral coefficients [26], using a 20-ms sliding analysis window with 10-ms frame shift. The resulting vector of audio features is denoted  $\mathbf{x}$ .

## 2.2. Building the articulatory model

For each of the two reference speakers, we first built an articulatory model using the general methodology originally proposed by [27], slightly adapted in [28] in order to process EMA articulatory data. In the present study, we consider the six following articulatory parameters (represented in Figure 1): *Jaw Height (JH)*, *Tongue Body (TB)*, *Tongue Dorsum (TD)*, *Tongue Tip (TT)*, *Lip Protusion (LP)*, *Lip Height (LH)*, *Velum (VL)* (for BY speaker only). Here “articulatory parameters” means that these parameters are interpretable in terms of articulatory control/function in speech production [29]. The core principle is to extract the latent dimension of tongue and lip movements after removing the contribution of the jaw, using a so-called “guided-PCA”. More precisely, the *JH* parameter (and the corresponding value for all the articulatory observations of the dataset) is defined as the first principal component (PC) of the jaw movement. The contribution of the jaw to the movement of the tongue is then estimated using a linear regression between *JH* values and the coordinated of the 3 EMA coils attached to the tongue. Once this contribution is estimated and removed, the parameters *TB* and *TD* are defined as the two first PCs of the joint movement of tongue dorsum and back. A linear regression between the *TB* and *TD* on one hand and the tongue tip coils on the other hand provides a residual movement of the tongue, freed from the contribution of the jaw, the tongue dorsum and the tongue back. The *TT* parameter is finally defined as the first PC of this residual movement. A similar procedure is used for extracting the *LP* and *LH* lip parameters and corresponding values for both datasets. For the BY2014 corpus, the *VL* parameter is simply defined as the first PC of the EMA coils attached to the velum. In summary, at the end of this guided PCA analysis, we have a linear transformation to go from a vector of EMA parameters  $\mathbf{y}$  to an articulatory vector  $\mathbf{a} = [JH, TB, TD, TT, LP, LH]$  (for the PB2007 dataset) or  $\mathbf{a} = [JH, TB, TD, TT, LP, LH, VL]$  for the BY2014 dataset, and vice versa. For convenience in the following, we denote by  $\mathbf{a}(\mathbf{x})$  the articulatory vector corresponding to the cepstral vector  $\mathbf{x}$ . Because this set of parameters is reduced in size compared to

EMA parameters, the articulatory information is “compressed” into a low-dimensional vector, which is appropriate for our later application of an articulatory constraint on the VAE latent space (which is also expected to be of low-dimension).

## 2.3. The VAE model

The seminal VAE model introduced in [14, 15] is defined by:

$$p_{\theta}(\mathbf{x}, \mathbf{z}) = p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z}), \quad (1)$$

where  $p(\mathbf{z})$ , the prior distribution of the latent vector  $\mathbf{z}$ , is a multivariate standard Gaussian distribution,  $p_{\theta}(\mathbf{x}|\mathbf{z})$  is the (conditional) likelihood function of the observed variable  $\mathbf{x}$ , and the dimension  $L$  of  $\mathbf{z}$  is (possibly much) lower than the dimension  $F$  of  $\mathbf{x}$ . The parameters of  $p_{\theta}(\mathbf{x}|\mathbf{z})$  are provided by a deep neural network (DNN), called the decoder network, that takes  $\mathbf{z}$  as input.  $\theta$  represents the parameters of this decoder network (e.g., the weights and biases of a multi-layer perceptron). In the present work,  $p_{\theta}(\mathbf{x}|\mathbf{z})$  is a conditional Gaussian distribution with diagonal covariance matrix.

Because the relationship between  $\mathbf{z}$  and  $\mathbf{x}$  is highly non-linear, the posterior distribution  $p_{\theta}(\mathbf{z}|\mathbf{x})$  is not analytically tractable. It is thus approximated with a parametric variational distribution  $q_{\phi}(\mathbf{z}|\mathbf{x})$ , a.k.a. the inference model, whose parameters are provided by another DNN (called the encoder network, with weights  $\phi$  and input  $\mathbf{x}$ ). A usual choice, that we follow here, is to set  $q_{\phi}(\mathbf{z}|\mathbf{x})$  as a conditional Gaussian distribution with diagonal covariance matrix. The parameters  $\{\theta, \phi\}$  are then jointly estimated by maximizing a lower bound of the data log-likelihood function, called the Variational Lower Bound (VLB), given by (for one single data vector):

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})], \quad (2)$$

and evaluated on a training dataset ( $D_{\text{KL}}$  denotes the Kullback-Leibler divergence). The left term of the VLB represents the reconstruction accuracy of the encoding-decoding process and the right term is a regularization term that ensures some degree of “disentanglement” of the latent vector entries [14]. Maximization of the VLB is done by combining stochastic gradient descent with sampling techniques.

## 2.4. The articulatory-regularised VAE (AR-VAE)

In the present work, each vector  $\mathbf{x}$  is a vector of Bark-scale cepstral coefficients, of dimension 18, extracted from the audio signal as described in Section 2.1. The dimension of  $\mathbf{z}$  will be specified later. To force the latent space of the VAE to fit the articulatory space, we added a third term to the above VLB using the same regularization technique as in [30] (itself inspired from [31]):

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\log p_{\theta}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q_{\phi}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})] + \alpha \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})} [\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x}))]. \quad (3)$$

The new regularization term  $\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x}))$  ensures that for each speech frame the first entries of the latent vector  $\mathbf{z}$  remain close to the corresponding entries in the vector of articulatory parameters  $\mathbf{a}(\mathbf{x})$  defined in Section 2.2. In our experiments, we implemented this regularization term with the mean squared error (MSE):

$$\mathcal{R}(\mathbf{z}, \mathbf{a}(\mathbf{x})) = \|\mathbf{z}_{1:N} - \mathbf{a}(\mathbf{x})\|^2, \quad (4)$$

where  $\mathbf{z}_{1:N}$  denotes the subvector made of the  $N$  first latent values, with  $N = 6$  or  $7$  depending on the dataset. This term can be

<sup>1</sup>available online at <http://doi.org/10.5281/zenodo.154083>

interpreted in statistical modeling terms as an additional Gaussian prior on  $\mathbf{z}_{1:N}$  with mean vector  $\mathbf{a}(\mathbf{x})$  and an arbitrary fixed variance. In practice, the two expectations in (3) are replaced with estimates based on Monte Carlo sampling of  $\mathbf{z}$  (using the well-known reparameterization trick, just as in the seminal VAE [14]). Finally,  $\alpha$  is a weighting factor monitoring the weight of the articulatory regularization term, and that will be varied in our experiments.

## 2.5. Implementation

We implemented the proposed articulatory-regularized VAE model with the following architecture: The encoder was composed of 4 fully connected hidden layers (256, 128, 64 and 32 neurons) and the decoder had the same, but reversed, composition. The size of the latent space was 12 (i.e., the size of the output layer of the encoder) when the model was trained on the PB2007 corpus (including 6 articulatory-regularized dimensions), and it was 14 when the model was trained on the BY2014 corpus (including 7 articulatory-regularized dimensions). Note that half of the  $\mathbf{z}$  entries are articulatory-constrained, hence they are forced to encode the information in cepstral vectors that is strongly correlated to the articulatory parameters, and the other half are let free to encode “everything else” (e.g., speech source information).

The hyperbolic tangent activation function was used for each hidden layer. Model training was done using back-propagation with Adam optimizer, on mini-batches of 32 observations (pairs of  $\mathbf{x}$  and  $\mathbf{a}(\mathbf{x})$  vectors). For each experience, the datasets were randomly partitioned with 80% of the data used for training and the remaining 20% used for testing. The implementation was done using the *PyTorch* toolkit [32].

## 3. Experiments

### 3.1. Model learning speed and accuracy

We first tested if the introduction of articulatory constraints could speed up the training process. To that purpose, for each of the two datasets, and for each value of  $\alpha$  taken in  $\{0, 0.1, 0.25, 0.5, 1\}$  we trained 10 AR-VAE models (with a different initialization each time) during 60 epochs (note that an AR-VAE with  $\alpha = 0$  is equivalent to a conventional VAE). At each epoch, we computed the reconstruction error defined as the MSE between the reconstructed and true cepstral coefficients, on the test set. For each value of  $\alpha$  (and for each dataset), we finally built a smooth version of the learning curve by averaging the reconstruction loss on the test set, over the 10 models. These learning curves are presented in Figure 2a. First, we observe that almost all AR-VAEs converge faster than the conventional VAEs (i.e. the blue dashed line is almost always above the other lines). Then, as summarized in Figure 2b, the best final performance is obtained with the proposed AR-VAE on both datasets (i.e. with  $\alpha = 1$  for PB2007 and with  $\alpha = 0.25$  for BY2014). Therefore, adding articulatory constraints does improve representation learning, both in terms of convergence speed and accuracy.

### 3.2. Robustness to noise

We then tested the performance of the proposed AR-VAE on a speech denoising task. To that purpose, a babble noise was added to each audio speech signal with different signal-to-noise ratios (SNR) (no noise, 10dB, 5dB and 0dB). Sequences of acoustic feature vectors (i.e. 18 Bark-scale cepstral coefficients)

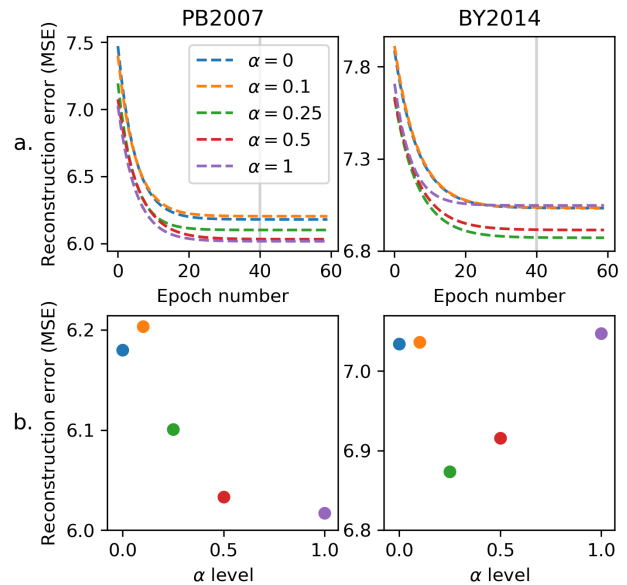


Figure 2: a.: Evolution of the reconstruction loss on the test set during training. For better visualization, each learning curve is fitted by an exponentially decreasing function b.: Final performance after convergence at epoch 40 on the test set.

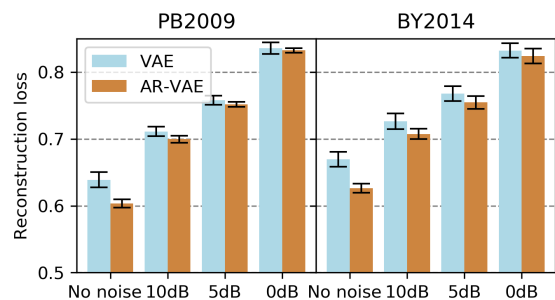


Figure 3: Reconstruction loss (mean  $\pm$  standard deviation) of the conventional VAE and the proposed AR-VAE on the test set for the speech denoising task.

icients) were extracted from the noisy audio signals with the same method used previously. For each dataset, both VAE and AR-VAE were trained to reconstruct the non-noisy version of each acoustic feature vector from its noisy counterpart. For the AR-VAE, we first compared different values of the  $\alpha$  parameter for this denoising task. For concision, we report here only the results with  $\alpha = 1$  which provided the best performance among the tested values. As for the first experiment, we trained 10 different VAEs/AR-VAEs (with a different initialization each time) and averaged the results over the 10 runs. The reconstruction error on the test dataset is shown in Figure 3. These results show that the proposed AR-VAE outperforms the conventional VAE on the denoising task, for all considered SNR. The performance difference is more pronounced for lower noise levels.

Since the absolute value of the reconstruction loss (which is in an arbitrary unit) is difficult to interpret, we then conducted further evaluations to assess the quality of the denoised speech. First, we evaluate its phonetic content. To that purpose, for each sentence of the test set, and for each considered SNR, we re-synthesized a speech signal using the LPC-

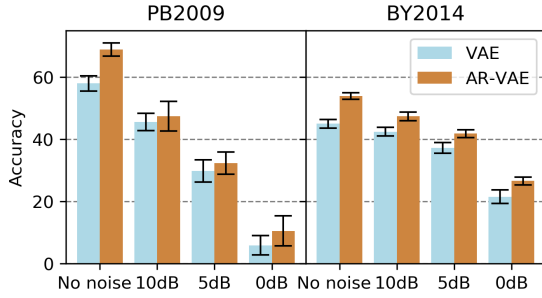


Figure 4: Accuracy (mean  $\pm$  standard deviation) of an HMM-based phonetic decoder when processing speech signals denoised by the conventional VAE and the proposed AR-VAE (with  $\alpha = 1$ ).

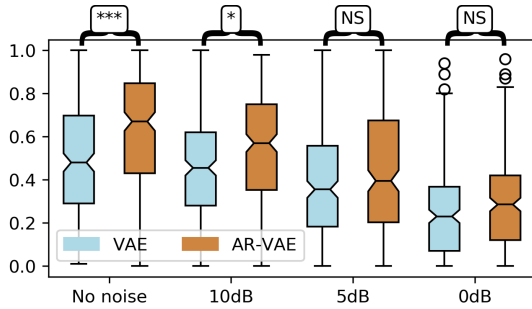


Figure 5: MUSHRA scores obtained for each level of noise and for the conventional VAE ( $\alpha = 0$ ) and the proposed AR-VAE (with  $\alpha = 1$ ). For the sake of clarity, we omit the anchor and reference scores. \*\*\* and \* are significant ( $p < 0.001$  and  $p < 0.05$ ), and NS are non-significant differences.

Net neural vocoder fed by the cepstral coefficients reconstructed by the VAE/AR-VAE, together with the original pitch parameters (period and correlation, extracted from the clean speech sound).<sup>2</sup> The resulting speech signals were sent to a Hidden Markov Model (HMM) based phonetic decoder, trained on the original (clean) speech signals of the training set (left-to-right 3 emitting states context-independent HMM-GMM, trained using the HTK toolkit and a standard procedure, no language model used). The decoding accuracy (which takes into account insertion and deletion errors) are presented in Figure 4. Again, the AR-VAE outperforms the VAE, by a large margin (up to 10%) when processing clean speech, and by a smaller one when processing noisy speech.

Finally, we assessed the perceptive quality of the reconstructed speech signal using a MUSHRA test during which participants had to rank a set of audio stimuli by their similarity with a reference sound on a scale from 0 to 100 [33]. We first randomly selected 20 short sentences from the BY2014 corpus (preferred to the PB2007 corpus because of the presence of data on the velum). For each sentence, we generated 6 audio stimuli: a low anchor built by adding babble noise to the original audio speech signal with an SNR of 0dB and re-synthesized (i.e. vocoded) with LPCNet, a hidden reference built by re-synthesizing the original signal with LPCNet, a third stimulus built by first encoding-decoding the original signal either with the conventional VAE or with the proposed AR-VAE, and then

<sup>2</sup>Several sound examples are available at <https://georges.ma/p/ar-vae>

synthesizing an audio signal with LPCNet, and four other stimuli built following the same principle but after having first added noise to the original signal with three levels of SNR (10dB, 5dB and 0dB). We recruited 23 native French speakers online via the Prolific Academic platform [34]. Results are reported in Figure 5. To assess the statistical significance of the difference between the MUSHRA scores, we first conducted a Kruskal-Wallis rank-sum test which showed a significant effect of the SNR factor ( $p < 0.05$ ). Then a post-hoc Dunn test validated a statistically significant increase of performance from VAE to AR-VAE for clean audio ( $p < 0.001$ ) and for noisy audio with SNR = 10dB, and showed no significant difference between the two models for SNR = 5dB and SNR = 0dB (i.e. very noisy inputs).

## 4. Discussion and perspectives

The two experiments reported in Section 3 suggest that articulatory constraints improve the learning of speech representations in a VAE. This study opens interesting perspectives for assessing the combined role of articulatory knowledge and auditory processes in the elaboration of internal representations of speech signals. While the efficiency of such articulatory-constrained VAE appears clearly in terms of spectrum reconstruction, it remains to explore how the corresponding representations are structured and whether they are able to integrate the concept of articulatory/motor invariance [35] inside phonetic representations. We will particularly focus on the way articulatory information in the present VAE could improve the representation of plosive place of articulation, known to depend on the availability of articulatory/motor information [36].

Another question concerns the way articulatory regularization provided by the articulatory data available for a given speaker could enable this speaker to better process the speech utterances of other speakers. For this aim, we will explore how the regularized VAE performs in a denoising experiment involving multiple speakers. We will also study other VAE architectures in which articulatory-acoustic data for the speaking agent and acoustic-only data for other agents could be learnt together in articulatory-constrained VAE variants.

Finally, the precise developmental schedule along which a child develops her production and perception skills and learns the sounds of her language and the corresponding articulatory trajectories will shed interesting light on some algorithmic choices that should be made in further developments of such articulatory/motor constrained acoustic VAEs.

## 5. Conclusion

In this paper, we show how articulatory knowledge can help construct internal representations of auditory speech stimuli, by applying an articulatory regularization to VAEs encoding speech features, enforcing them to adopt mixed articulatory-acoustic representations in their latent space. The additional term in the training loss function enforcing a part of the latent space to follow articulatory parameters appears to improve learning efficiency, both in terms of learning speed and accuracy, and also improve reconstruction performance in a denoising VAE.

## 6. Acknowledgement

This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003). The authors would like to thank Pierre Badin and Julien Diard for fruitful discussions.

## 7. References

- [1] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.
- [2] J.-L. Schwartz, A. Basirat, L. Ménard, and M. Sato, "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 336–354, 2012.
- [3] J. I. Skipper, J. T. Devlin, and D. R. Lametti, "The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception," *Brain and Language*, vol. 164, pp. 77–105, 2017.
- [4] J. F. Houde and S. S. Nagarajan, "Speech production as state feedback control," *Frontiers in Human Neuroscience*, vol. 5, 2011.
- [5] F. H. Guenther and T. Vladusich, "A neural theory of speech acquisition and production," *Journal of Neurolinguistics*, vol. 25, no. 5, pp. 408–422, 2012.
- [6] R. C. Rose, J. Schroeter, and M. M. Sondhi, "The potential role of speech production models in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 99, no. 3, pp. 1699–1709, 1996.
- [7] C. Castellini, L. Badino, G. Metta, G. Sandini, M. Tavella, M. Grimaldi, and L. Fadiga, "The use of phonetic motor invariants can improve automatic phoneme discrimination," *PLoS ONE*, vol. 6, no. 9, p. e24055, sep 2011.
- [8] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.
- [9] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory information for noise robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 1913–1924, 2011.
- [10] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.
- [11] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.
- [12] R. Laurent, M.-L. Barnaud, J.-L. Schwartz, P. Bessière, and J. Diard, "The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception," *Psychological Review*, vol. 124, no. 5, pp. 572–602, 2017.
- [13] T. Hueber, E. Tatulli, L. Girin, and J.-L. Schwartz, "Evaluating the potential gain of auditory and audiovisual speech predictive coding using deep learning," *Neural Computation*, pp. 596–625, 2019.
- [14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. of ICLR*, Banff, Canada, 2014.
- [15] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic back-propagation and approximate inference in deep generative models," in *Proc. of ICML*, Beijing, China, 2014.
- [16] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," in *Proc. of Interspeech*, San Francisco, CA, 2016.
- [17] W. N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. of Interspeech*, Stockholm, Sweden, 2017, pp. 1273–1277.
- [18] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. of APSIPA*, Jeju, Korea, 2016.
- [19] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. of ICASSP*, Calgary, Canada, 2018.
- [20] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proc. of Interspeech*, Hyderabad, India, 2018.
- [21] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multi-channel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. of ICASSP*, Brighton, UK, 2019.
- [22] R. Turrisi, R. Tavarone, and L. Badino, "Improving generalization of vocal tract feature reconstruction: From augmented acoustic inversion to articulatory feature reconstruction without articulatory data," in *Proc. of IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp. 159–166.
- [23] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology Yearbook*, vol. 3, pp. 219–252, 1986.
- [24] P. Saha and S. Fels, "Learning Joint Articulatory-Acoustic Representations with Normalizing Flows," in *Proc. of Interspeech*, oct 2020, pp. 3196–3200.
- [25] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. of ICASSP*, Brighton, UK, 2019, pp. 5891–5895.
- [26] M. Schroeder, B. Atal, and J. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of Speech Communication Research*, L. Björn, S. Öhman, and G. Fant, Eds. Academic Press, 1979, pp. 217–229.
- [27] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," *Speech Production and Speech Modelling*, pp. 131–149, 1990.
- [28] A. Serrurier, P. Badin, A. Barney, L.-J. Boë, and C. Savariaux, "The tongue in speech and feeding: Comparative articulatory modelling," *Journal of Phonetics*, vol. 40, no. 6, pp. 745–763, nov 2012.
- [29] S. Maeda and K. Honda, "From EMG to Formant Patterns of Vowels: The Implication of Vowel Spaces," *Phonetica*, vol. 51, no. 1-3, pp. 17–29, 1994.
- [30] F. Roche, T. Hueber, M. Garnier, S. Limier, and L. Girin, "Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder," *Transactions of the International Society for Music Information Retrieval*, vol. accepted - pending publication, 2021.
- [31] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Generative timbre spaces with variational audio synthesis," in *Proc. of DAFx*, Aveiro, Portugal, 2018.
- [32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. of NeurIPS*, Vancouver, Canada, 2019, pp. 8024–8035.
- [33] ITU, "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Tech. Rep. ITU-R BS.1534-3, October 2015.
- [34] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.
- [35] A. Liberman, F. Cooper, D. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological review*, vol. 74, no. 6, pp. 431–461, 1967.
- [36] M.-L. Barnaud, J. Diard, P. Bessière, and J.-L. Schwartz, "COSMO SylPhon: A bayesian perceptuo-motor model to assess phonological learning," in *Proc. of Interspeech*, Hyderabad, India, 2018, pp. 3786–3790.