A DNN Based Speech Enhancement Approach to Noise Robust Acoustic-to-Articulatory Inversion

Abdolreza Sabzi Shahrebabaki, Sabato Marco Siniscalchi, Giampiero Salvi, Torbjørn Svendsen
Department of Electronic Systems, NTNU
Email: {abdolreza.sabzi, marco.siniscalchi, giampiero.salvi, torbjorn.svendsen}@ntnu.no

Abstract—**In this work, we investigate the problem of speaker independent acoustic-to-articulatory inversion (AAI) in noisy condition within the deep neural network (DNN) framework. We claim that DNN vector-to-vector regression for speech enhancement (DNN-SE) can play a key role in AAI when used in a front-end stage to enhance speech features before AAI back-end processing. Our claim contrasts recent literature reporting a drop in AAI accuracy on MMSE enhanced data and thereby sheds some light on the opportunities offered by DNN-SE in robust speech applications. We have also tested single- and multi-task training strategies of the DNN-SE block and experimentally found the latter to be beneficial to AAI. Moreover, DNN-SE coupled with an AAI deep system tested on enhanced speech can outperform a multi-condition AAI deep system tested on noisy speech. We assess our approach on the Haskins corpus using the Pearson's correlation coefficient (PCC). A 15% relative PCC improvement is observed over a multi-condition AAI system at 0dB signal-to-noise ratio (SNR). Our approach also compares favorably against using a conventional DSP approach, namely MMSE with IMCRA, in the front-end stage.**

Index Terms: Acoustic-to-articulatory inversion, DNN, speech enhancement

I. Introduction

The use of articulatory parameters has attracted increasing interest in the speech field because it has been proven beneficial in applications such as low bit rate coding [1], automatic speech recognition (ASR) [2], [3], [4], speech synthesis [5], and speech therapy [6], [7]. The articulators' movements can be measured and parameterized using different techniques; for instance, real-time magnetic resonance imaging (rt-MRI) [8], X-ray microbeam [9], electromagnetic articulography (EMA) [10], and ultrasound [11]. Nevertheless, obtaining articulatory measurements is not practical in real-world applications. The alternative is to estimate the articulatory parameters from speech recordings, which are more easily accessible, with a process called acoustic-to-articulatory inversion (AAI). Unfortunately, the AAI problem is highly non-linear and non-unique [12], [3], which means that different articulator configurations can produce the same sound. In addition, coarticulation e,g, [12], i.e., the influence of adjacent phonemes on the articulators' movement, makes the AAI problem even harder. Finally, related to the topic of this paper, the aforementioned applications require solutions that work independently of the speaker, and, possibly, in noisy acoustic conditions.

There exist several techniques to address the AAI problem, for example, search-based algorithms in the joint codebook of the acoustic-articulatory space [13], [14], non-parametric and parametric statistical methods such as support vector regression (SVR) [15], joint acoustic-articulatory distribution by utilizing Gaussian mixture models (GMMs) [16], hidden Markov models (HMMs) [17], mixture density networks (MDNs) [18], deep neural networks (DNNs) [19], and recurrent neural networks (RNNs) [20], [21]. However, the great majority of those works deals with clean conditions only.

There are a few works using synthetically generated articulatory trajectories for robust ASR [12], [4], where an AAI system was trained on multi-condition noisy data. However, we found only a single study about real articulatory measurements in noisy conditions [22], which claims that is more beneficial to train the AAI system with multi-condition noisy data rather than performing speech enhancement as a pre-processing step to an AAI system trained on clean data. That result could be explained noticing that a digital signal processing (DSP) approach to speech enhancement based on minimum mean square error (MMSE) [23] was used. Therefore, improved perceptual quality (e.g. higher PESQ) is attained rather than reduced SNR. Furthermore, artifacts and attenuation introduced in the signal by DSP enhancement may create a mismatch between training and testing conditions, which lead to a performance degradation.

However, DNN-based approaches to speech enhancement, dubbed DNN-SE, can outperform state-of-the-art conventional DSP ones [24]. Moreover, employing a DNN-SE preprocessing step before the target speech applications has been shown to be beneficial [25], [26]. We thus propose to address the AAI problem in noisy conditions by coupling a DNN-SE approach as a front-end pre-processing stage with the back-end AAI. In particular, we use the DNN-based vector-to-vector regression approach to speech enhancement that was demonstrated to be highly robust [24]. Experiments are carried out on the Haskins production rate comparison database (HPRC) [27]. AAI performance is assessed using the Pearson's correlation coefficient, $PCC = cov(y, \hat{y})/(\sigma_y \sigma_{\hat{y}})$.

To have a comprehensive assessment, we compare and contrast our approach with an AAI system paired with a conventional speech enhancement approach, based on MMSE with improved minima controlled recursive averaging (IM-

S.M. Siniscalchi is also with Department of Computer Engineering, Kore University of Enna. G. Salvi is also with KTH Royal Institute of Technology, Dept. of Electrical Engineering and Computer Science,

This work was supported by the Research Council of Norway AULUS, and the NTNU ArtiFutt projects. Prof. Siniscalchi is supported by PRIN2017 n. 2017JNKCYZ_002 project.

CRA) [28]. We investigate different DNN-SE configurations: speaker-independent training with matched and mismatched test and training speakers; single-task training, where the DNN-SE estimates only log power spectra (LPS) later converted to MFCCs to be used as an AAI input; and multi-task training, where both MFCCs and LPS are estimated by DNN-SE. Multi-task training is shown to not only be beneficial for speech enhancement but also for the AAI.

II. Deep neural models

A. DNN based acoustic-to-articulatory inversion

A fully-connected, feed-foward DNN is built to accomplish AAI. The DNN-AAI system is speaker-independent (SI), i.e., leave-one-speaker-out (LOSO) trained. The input features are Mel frequency cepstral coefficients (MFCCs), which attained the highest AAI accuracy in SI conditions [29] by removing the higher order cepstral coefficients, that describe the spectral fine structure. Due to the smooth nature of articulatory trajectories and co-articulation effects, the temporal context at the input acoustic layer should be long enough [3] to capture useful information with respect to the output trajectories. M_{AAI} frames around the current input frame are thus used.

B. DNN based speech enhancement

The DNN-SE system has three ReLU-based non-linear layers. The input feature vectors are globally mean and variance normalized LPS, extracted as the log squared magnitude of the short-time Fourier transform of the signal. Test set feature vectors are normalized using mean and variance information obtained on the training data. The noisy phase information is not processed during the enhancement process and is only used to reconstruct the speech signal, as in [24]. M_{SE} previous and future frames around the current frame are used at the DNN-SE input layer. The temporal context M_{SE} is shorter than M_{AAI}. This shorter context is congruous with the non-stationary property of noises, enabling the network to have a better estimation of the short-time noise spectrum to be suppressed. The DNN-SE system is trained in two different ways. In single-task training the network predicts clean LPS frames given noisy LPS frames as input. In multi-task training, the network predicts at the same time the clean LPS frames and the MFCCs.

III. Corpora

This work is concerned with AAI in noisy condition. We thus artificially added noise from the Aurora 2 corpus [30] to the Haskins production rate comparison database (HPRC) [27] speech waveforms and obtained multi-condition speech data paired with articulatory trajectory measurements. Aurora 2 audio recordings are from eight different environments, namely airport, babble, car, exhibition, restaurant, street, subway and train. The sampling rate is 8kHz. Some of these noises include non-stationary parts, e.g. the street and the airport, and some of them are fairly stationary like car and exhibition.

The HPRC database is a multi-speaker speech database that contains synchronous EMA recordings. The original sampling rate of the audio recordings is 44.1kHz and the EMA parameters are recorded at rate of 100 Hz. Eight native American English speakers (half male, half female) were asked to read 720 different sentences from the IEEE sentence list [31] in normal and fast speaking style. We used sensor measurements from the tongue rear (TR), tongue blade (TB), tongue tip (TT), lower lip (LL), upper lip (UL) and jaw (JAW) in the X and Z directions, denoting sensor movements from the posterior to the anterior and from the inferior to the superior, respectively. We converted the articulators' measurements to tract variables (TVs) by the geometrical transformations defined in [32], [33]. The TVs represent the constriction degree and location of the articulators producing the speech. In contrast with the EMA measurements TVs are calculated relatively and suffer less from non-uniqueness [34]. The HPRC audio data was downsampled to 8kHz to match the Aurora recordings.

IV. Experimental setup

A. Multi-condition data

We use the normal speaking style utterance from the HPRC corpus. For each speaker, 80% of the utterances are used for training, 10% for for validation, and 10% for testing. Noise is artificially added to the clean speech signals to simulate noisy working conditions. All eight available noises from the Aurora 2 corpus were used to corrupt the clean HPRC speech waveforms, using five signal-to-noise ratio (SNR) levels, from 0dB to 20dB with a step size of 5dB. Therefore, the noisy data size is 40 times bigger than that of the clean data size, which covers roughly around four hours. That is, we have generated 160 hours of multi-condition data, including clean data.

B. Feature representation

For the AAI system, as it is described previously, the input acoustic features are (MFCCs) and the outputs are tract variables. The input and output of the AAI system are z-score normalized per utterance which is suggested by [22]. The DNN-SE system uses noisy LPS as input and either clean LPS or clean LPS and MFCCs as outputs. The input and output LPS are normalized with the global mean and variance. The TVs' rate is 100 Hz, so MFCCs and LPSs were extracted with a 20ms window and a 10 ms shift. The MFCCs were 13 dimensional vectors computed over 23 Mel filterbanks.

V. Experiments and results

Experimental evidence to assess the viability and effectiveness of the proposed approach is given in the next sections.

A. AAI trained with clean data

As indicated in Section II, the speaker-independent AAI model is fed with MFCC vectors and estimates TVs at its output. The temporal context M_{AAI} is set equal to 8 frames, which spans a $340\mathrm{ms}$ segment. In clean conditions, several experiments have been carried out to select the number of hidden layers and hidden nodes per layer. We experimented with the following configurations: $[100, 300, 500, 1000]$ nodes, and $[2, 3, 4, 5]$ hidden layers. The PCC value is given in

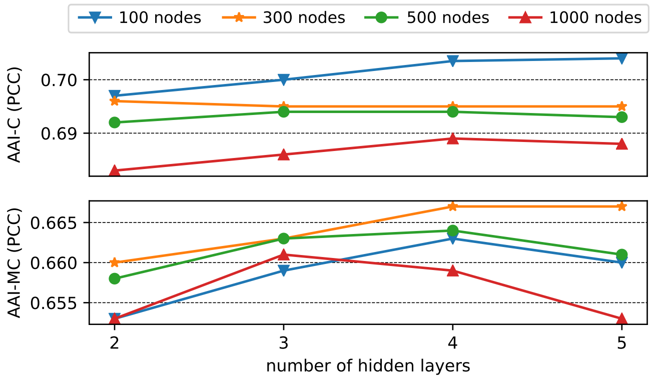

Fig. 1. Average PCC performance vs AAI DNN parameters with matched training and test data: clean data (top panel) and multi-condition data (bottom panel)

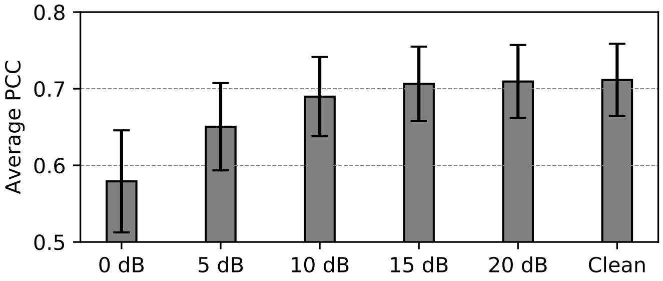

Fig. 2. Average PCC for AAI-MC on noisy data in terms of SNR levels.

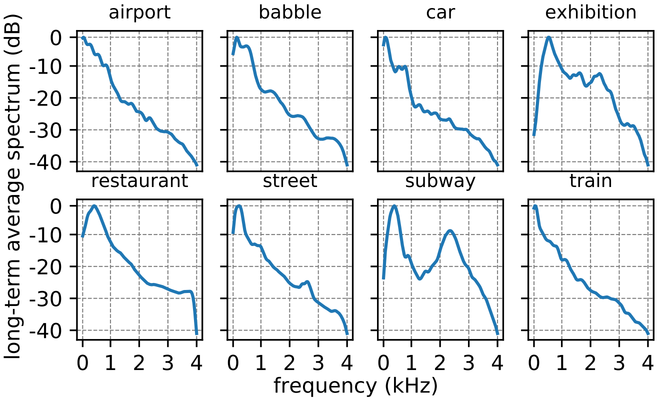

Fig. 4. Long-term average spectrum of different noise types in Aurora 2.

the upper panel in Figure 1. The results show us the best performing system is a 5 hidden layer network with 100 nodes in each layer. As the amount of available data is limited it is reasonable that increasing the number of parameters beyond a certain point will not improve the performance. The speaker-independent AAI system trained on clean data is referred to as AAI-C.

B. AAI trained with multi-condition data

Following the procedure highlighted in Section V-A, we tune the speaker-independent AAI system trained on multi-condition data. Examining the lower panel in Figure 1, we chose a configuration with 4 hidden layers of 300 nodes. This speaker-independent AAI system is referred to as AAI-MC. After parameter tuning, we focus on exploring (i) effect of the SNR level, and (ii) effect of the noise type on the AAI performance in noisy conditions. Results for various SNR levels are depicted in Figure 2, where we observe near constant PCCs for SNRs ≥ 15dB. High standard deviations in the average PCC is due to several factors, e.g. LOSO cross-validation, different variation range for each of the TVs and the effect of various noise types and the SNR levels. The PCC averages across different speakers and SNRs indicate the influence of different noise types on the performance of AAI-MC and AAI-C systems. In Figure 3, it can be seen that ´exhibition´ and ´subway´ noises have the most adverse effect on performance, whereas ´car´ and ´train have less effect on the performance. A visual inspection of the long-term average power spectrum of different noise types, given in Fig. 4, hints that noises with considerable energy in frequency band 1kHz to 3kHz cause a more severe degradation of the AAI performance. For clean data, AAI-MC performs slightly better than the AAI-C, which can be explained by having more training data for almost clean conditions (SNR $\geq$ 20dB), i.e., a data-augmentation effect.

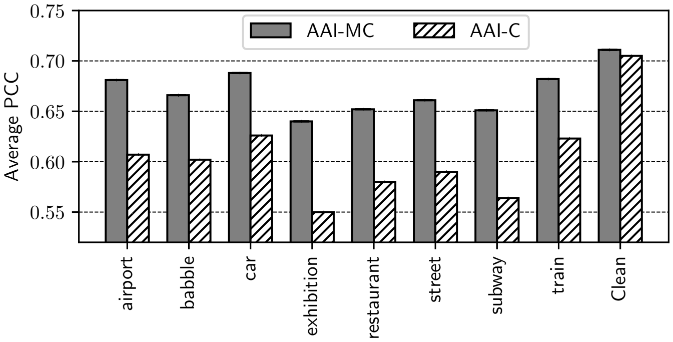

Fig. 3. Average PCC for AAI-C and AAI-MC on different noise types.

C. DNN based speech enhancement

Speech enhancement experiments are carried out in different conditions, namely (i) matched (DNN-SE-Match-Spk) versus mismatched (DNN-SE-MisMatched-Spk) speaker in training and testing, (ii) single-task (ST) versus multi-task (MT) training. DNN-SE-Match-Spk is trained on many speakers, so it is broadly speaking speaker-independent, but test speakers are seen during training too. To remove such a limiting factor in a real scenario, a DNN-SE-MisMatch-Spk system is built. A combination of different aspects provides us with more insights for further investigation. Table I shows the average of the perceptual evaluation of speech quality (PESQ) [35] for different systems and testing conditions. From Table I, we can see that DSP-SE improves, as expected, the average PESQ, most for intermediate noise levels. However, DNN-SE outperforms DSP-SE in line with [24], but most importantly

TABLE I
PERFORMANCE COMPARISON OF SINGLE-TASK AND MULTI-TASK DNN BASED ENHANCEMENT SYSTEMS IN MATCHED AND MISMATCHED TEST SPEAKER CASES.

SNR	Noisy	DSP-SE	DNN-SE-Match-Spk		DNN-SE-MisMatch-Spk	
			ST	MT	ST	MT
0 dB	1.51	1.700	2.554	2.653	2.365	2.528
5 dB	1.75	2.077	2.767	2.873	2.544	2.729
10 dB	2.06	2.533	2.955	3.069	2.702	2.907
15 dB	2.47	2.950	3.104	3.224	2.828	3.048
20 dB	2.97	3.316	3.205	3.333	2.919	3.148

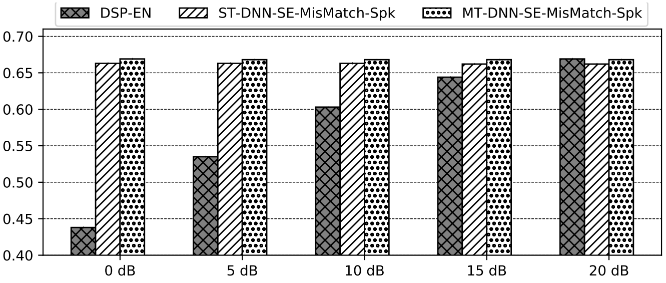

Fig. 5. PCC of AAI-C on enhanced speech at different SNRs.

TABLE II
PCC FOR AAI-C AND AAI-MC ON CLEAN, MULTI-CONDITION AND ENHANCED DATA.

Test data	Enhancement	AAI-C	AAI-MC
Clean	None	0.705	0.710
Multi-Cond	None	0.595	0.665
Multi-Cond	DSP-SE	0.568	0.620
Multi-Cond	MT-DNN-SE-Match-Spk	0.699	0.711
Multi-Cond	ST-DNN-SE-Match-Spk	0.689	0.702
Multi-Cond	MT-DNN-SE-MisMatch-Spk	0.670	0.711
Multi-Cond	ST-DNN-SE-MisMatch-Spk	0.662	0.693

it works much better than DSP-SE in low SNRs. DNN-SE-Match-Spk systems can be trained using single-task and multi-task strategy. Multi-task gives an increment of 0.1 over the single-task approach in terms of average PESQ. The MFCC output acts as a regularizer for the LPS output, as hinted by improvements in the average PESQ, and vice versa by looking at the PCC values for AAI in Table II. DNN-SE-MisMatch-Spk shows a drop of $\approx$ 0.2 in PESQ value in single-task training and $\approx$ 0.15 PESQ in multi-task training with respect to the DNN-SE-Match-Spk counterparts.

D. AAI with enhanced speech data

The first step is to verify whether DNN-based speech enhancement is useful and can boost the accuracy of the AAI-C system. To this end, we use the DNN-SE-MisMatch-Spk, which is more suitable for real-world speech applications. AAI-C tested on data enhanced by MT-DNN-SE-MisMatch-Spk performs slightly better than AAI-MC tested on multi-condition data. It can be argued that the improvement comes from an increase of the neural parameters caused by coupling two deep models. Yet, the DNN-SE and AAI-C deep model were independently trained on different data, and our solution allows to use an off-the-shelf AAI-C system avoiding training a system from scratch, an aspect that should not be overlooked in a production pipeline of a real complex system. It should be recalled that [22] reported DSP-SE to cause a drop in the AAI performance. We therefore further compare DSP-SE and DNN-SE-MisMatch-Spk effects on AAI-C. In Table II, we see that DSP-SE coupled with AAI-C causes a 0.14 drop in the PCC compared to MT-DNN-SE-MisMatch-Spk coupled with AAI-C. Most importantly, AAI-C with DSP-SE enhanced data reduces PCC by 4.5% relative compared to AAI-C tested on noisy data. That result confirms [22], and it could be explained by possible signal distortions introduced by DSP-SE. The DNN-SE method does not cause a drop in AAI performance, shedding new light on the use of DNN-based front-end approaches for speech applications. For the sake of completeness, Table II shows experimental results with multi-task (MT) and single-task (ST) training strategies, in matched and mismatched speaker scenarios. MT-DNN-SE methods outpeform ST-DNN-SE counterparts, whereas a drop in PCC is observed when moving from matched to mismatched training/testing speakers. Finally, we also provide results using AAI-MC on clean, noisy and enhanced data. Interestingly, our proposal improves also AAI-MC performance A detailed comparison in terms of SNR values of the AAI-C on DSP-SE and DNN-SE-MisMatch-Spk enhacned data is shown in Figure 5. Enhancement with DNN-SE-MisMatch-Spk always gives a better PCC in low SNR conditions; moreover, DNN-SE-MisMatch-Spk and DSP-SE lead to similar PCC only in very high SNR. At 0dB, from Figures 3 and 5, we see that AAI-MC attains a PCC of 0.579, and AAI-C on DNN-SE enhanced data attains a PCC of 0.67, which accounts for a 15% relative improvement in favor of the proposed DNN-SE based AAI-C approach.

VI. CONCLUSION

We have investigated speaker-independent AAI for noisy speech, showing that DNN-based speech enhancement can boost an AAI-C system trained on clean data. Good improvement was observed for an AAI-MC system trained on multi-condition data. Moreover, the AAI-C system showed no drop in performance when moving from clean to enhanced data at testing time and matched speaker. In mismatched-speaker scenarios, the AAI-C system performed better than the AAI-MC system on multi-condition data, which clearly demonstrates the effectiveness of the proposed speech enhancement pre-processing with deep models. In future work, we will investigate advanced methods to cope with matched- and mismatched-speaker scenarios at a DNN-SE level. Moreover, joint training of the enhancement and inversion could also be beneficial to both tasks.

References

[1] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," *Advances in Speech Signal Processing*, pp. 231–267, 1992.

[2] J. Frankel and S. King, "ASR-articulatory speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[3] V. Mitra, "Articulatory information for robust speech recognition," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2010.

[4] 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, Sep. 2011.

[5] K. Richmond and S. King, "Smooth talking: Articulatory join costs for unit selection," in *ICASSP*, 2016, pp. 5150–5154.

[6] D. W. Massaro, S. Bigler, T. Chen, M. Perlman, and S. Ouni, "Pronunciation training: the role of eye and ear," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[7] S. Fagel and K. Madany, "A 3-d virtual head as a tool for speech therapy for children," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[8] S. Narayanan, K. N. S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[9] J. R. Westbury, G. Turner, and J. Dembowski, "X-ray microbeam speech production database user's handbook," *University of Wisconsin*, 1994.

[10] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[11] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "Dnn-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8.

[12] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[13] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[14] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[15] A. Toutios and K. Margaritis, "A support vector approach to the acoustic-to-articulatory mapping," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[17] T. Hueber, A. Ben Youssef, G. Bailly, P. Badin, and F. Elisei, "Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory HMM for pronunciation training," in *Interspeech*, 2012, pp. 783–786.

[18] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[19] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *ICASSP*, 2015, pp. 4450–4454.

[21] A. S. Shahrebabaki, N. Olfati, A. S. Imran, S. M. Siniscalchi, and T. Svendsen, "A Phonetic-Level Analysis of Different Input Features for Articulatory Inversion," in *Interspeech*, 2019, pp. 3775–3779.

[22] N. Seneviratne, G. Sivaraman, V. Mitra, and C. Espy-Wilson, "Noise robust acoustic to articulatory speech inversion," in *Proc. Interspeech 2018*, 2018, pp. 3137–3141. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1509

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[25] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement," in *Proc. Interspeech 2019*, 2019, pp. 1691–1695. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1811

[26] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.

[27] M. Tiede, C. Y. Espy-Wilson, D. G. V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[28] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[29] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[30] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[31] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[32] A. Ji, "Speaker independent acoustic-to-articulatory inversion," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2014.

[33] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019. [Online]. Available: https://doi.org/10.1121/1.5116130

[34] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, no. 1, pp. 19 – 48, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167639394900558

[35] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.