INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Raw Speech-to-Articulatory Inversion by Temporal Filtering and Decimation

Abdolreza Sabzi Shahrebabaki[1], *Sabato Marco Siniscalchi*[1,2,3], *Torbjørn Svendsen*[1]

[1]Department of Electronic Systems, NTNU, Norway
[2]Department of Computer Engineering, Kore University of Enna, Italy
[3]Georgia Institute of Technology, USA

{abdolreza.sabzi,torbjorn.svendsen}@ntnu.no, marco.siniscalchi@unikore.it

Abstract

We propose a novel sequence-to-sequence acoustic-to-articulatory inversion (AAI) neural architecture in the temporal waveform domain. In contrast to traditional AAI approaches that leverage hand-crafted short-time spectral features obtained from the windowed signal, such as LSFs, or MFCCs, our solution directly process the input speech signal in the time domain, avoiding any intermediate signal transformation, using a cascade of 1D convolutional filters in a deep model. The time-rate synchronization between raw speech signal and the articulatory signal is obtained through a decimation process that acts upon each convolution step. Decimation in time thus avoids degradation phenomena observed in the conventional AAI procedure, caused by the need of framing the speech signal to produce a feature sequence that perfectly matches the articulatory data rate. Experimental evidence on the "Haskins Production Rate Comparison" corpus demonstrates the effectiveness of the proposed solution, which outperforms a conventional state-of-the-art AAI system leveraging MFCCs with an 20% relative improvement in terms of Pearson correlation coefficient (PCC) in mismatched speaking rate conditions. Finally, the proposed approach attains the same accuracy as the conventional AAI solution in the typical matched speaking rate condition.

Index Terms: Acoustic-to-articulatory inversion, raw speech modelling, 1D-convolution, temporal convolutional network (TCN)

1. Introduction

Acoustic-to-articulatory inversion (AAI) refers to the problem of estimating the parameters that describe the movement of the articulators from the uttered speech. In recent years, AAI has attracted increasing attention because of its potential applications in speech processing. Examples include low bit rate coding [1], automatic speech recognition (ASR) [2, 3, 4], speech synthesis [5, 6], computer aided pronunciation training (CAPT) [7, 8], depression detection from speech [9, 10], and speech therapy [11, 12]. Several regression-based methods were devised to to deal with the AAI problem before the deep learning breakthrough. For example, non-parametric and parametric statistical methods, such as support vector regression (SVR) [13], joint acoustic-articulatory distribution by utilizing Gaussian mixture models (GMMs) [14], hidden Markov models (HMMs) [7], mixture density networks (MDNs) [15]. State-of-the-art approaches leverage sequence-to-sequence deep models, for example, recurrent neural networks (RNNs) in [16, 17, 18, 4, 19].

Interestingly, deep and non-deep methods focused mainly on properly tackling the high non-linearity and non-uniqueness issues in the AAI task. The speech representation commonly adopted was in the short-time frequency domain, e.g., Line Spectral Frequencies (LSFs) [20], Perceptual Linear Predictive coding (PLP) [21] and Mel-Frequency Cepstral Coefficients (MFCCs)[22]. Filter-Bank Energies (FBEs) from STRAIGHT spectra [23] have also been employed as the input of the AAI system [18], which uses a parametric modelling of the speech spectrum, and the human auditory system. Those hand-crafted speech features have been adopted due to their success in different speech processing areas, for instance, LSFs was useful in speech coding [24], and voice conversion [25], FBEs and MFCCs were widely adopted with success in speech recognition, speaker recognition [26], and voice conversion [27]. The first required step in extracting those features is the windowing of the speech signal in the time domain in the hope of satisfying the requirements of stationarity posed by the Fourier transform. However, the windowing typically has a fixed window duration and shift. A fixed analysis window and consequent constant frame rate are not optimal settings for modeling the different characteristics of different parts of speech signal [28]. In fact, non-stationary parts, such as plosives and transient speech, have shorter duration compared to the stationary parts (e.g., vowels). Such a deficiency causes a performance degradation in the final speech application, especially when the speaking rate (SR) becomes slower or faster [29] compared to a normal speaking rate: Changes in SR affect both dynamic and static properties of speech. The former are related to the duration of phonemes and their transient phase. The latter are related to the distortion in the spectrum: "This distortion may be caused by the unusual movement of articulators particularly when dealing with co-articulations" [29]. In addition, there are several works in speech applications [30, 31, 32] argueing that for particular tasks using fixed filterbanks is not the optimal choice.

The above mentioned issues motivated us to leverage 1D convolutional layers and decimation to extract suitable features for the AAI task directly in the tempora domain. The proposed solution will be presented in Section 2, where the key features to avoid the degradation phenomena are discussed. In Section 4, the experimental evidence will be reported, which clearly demonstrates the advantages of the proposed approach over conventional approaches. In particular, we demonstrate comparable results with state-of-the-art conventional AAI solution when speech features and articulatory features are synchronous. Moreover, our solution based on the raw speech waveform for articulatory inversion outperforms the conventional state-of-the-art AAI system leveraging MFCCs by an 20% relative improvement in terms of Pearson correlation coefficient (PCC) in mismatched speaking rate.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1429

2. Proposed Method

In the proposed method, the raw waveform is directly utilized to accomplish the AAI task. To deal with the mismatch in sampling rate between the acoustic speech signal and articulatory measurements, - the sampling rate of speech signal is much higher than that of the articulatory signal, a multi-stages decimation procedure is employed. Decimation can be accomplished by pooling layers - we use max-pooling layers, or leveraging the stride operation in the convolutional layers - samples are skipped while sliding the convolutional filters over the input. In this work, we employ both max-pooling layers and strides to decimate the input signal and reduce its rate to that of the articulatory signal, namely 100 Hz. The decimation is done gradually in several stages, which allows to cover a much bigger temporal span compared to that of hand-crafted features, which is limited to the frame length. Furthermore, using the max-pooling operation with overlaps provides a non-uniform downsampling of the signal that preserves the required information for the AAI task from the relevant region of speech. This is in contrast with the fixed and uniform downsampling factor needed to match the articulatory rate when extracting handcrafted speech features.

After having the decimated the input to match the target articulatory rate, a temporal convolutional network (TCN) [33, 34] is employed to captures the dynamics in the speech signal, which are beneficial for the estimation of articulators' movements. TCNs use hierarchy of temporal causal convolutions to capture short and long range patterns from the input signal leveraging upon dilated convolutions. The filter size k and dilation factor d affect the receptive field of a TCN. The receptive field of the TCN can be increased by choosing larger filter size, and augmenting the dilation factor so that the receptive field can cover the temporal length of $(k-1)d$. One of the TCN's key strengths is the possibility of parallelizing the operations in contrast to RNNs. Finally, the TCN output is fed into a 1D convolutional layer followed by a time distributed fully connected layer to estimate the articulatory information.

3. Experimental Setup

3.1. Database

The EMA method is one of the most used techniques for the recording of articulatory data, which also allows for simultaneous recording of the speech signal. One of the available databases with EMA recording is the "Haskins Production Rate Comparison"(HPRC) [35], which covers material from eight native American English speakers, namely four female (F1-F4), and four male (M1-M4) speakers. There are 720 sentences available in this database with the normal and fast Speaking Rate (SR). For some of the normal speaking utterances, there are repetitions available. The amount of data for each speaking rate (SR) is shown in Table 1, where ''N1'', ''N2'' and ''F1'' represent the normal SR; repetition of some of the sentences with the normal SR; and fast SR, respectively.

Speech waveforms are sampled at rate of 44.1 kHz, and the synchronously recorded EMA data are sampled at 100 Hz. EMA data is measured from eight sensors capturing information about the tongue rear or dorsum (TR), tongue blade (TB), tongue tip (TT), upper and lower lip (UL and LL), mouth left (ML), jaw or lower incisors (JAW) and jaw left (JAWL). The articulatory movements are measured in the midsagittal plane in X, Y and Z direction, which denote movements of articulators from posterior to anterior, right to left and inferior to superior, respectively. In this work, we used the X and Z directions of TR, TB, TT, UL, LL and JAW for the speaker dependent AAI.

Table 1: *Available amount of data in HPRC database.*

SR	NO. utterances	Amount of data (minutes)
N1	5756	∼ 244
N2	1379	∼ 55
F1	5735	∼ 173

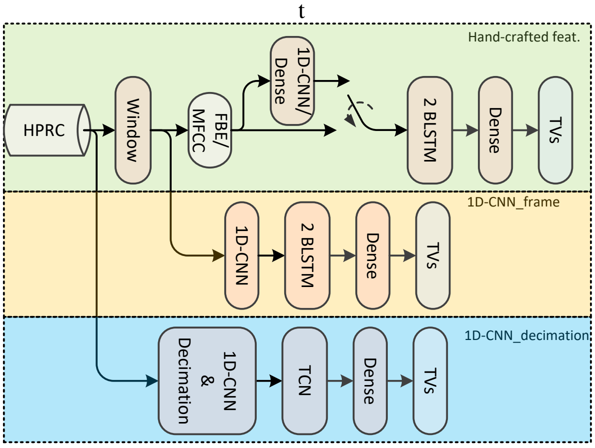

Figure 1: *The state-of-the-art S2S-AAI systems, employing (top) hand-crafted features, (middle) extracted features from speech frames by 1D-CNN, (bottom) extracted features from the whole speech sequence by 1D-CNN and decimation layers.*

The speech waveforms are downsampled to 16 kHz for performing AAI. For each of the fast and normal speaking rates, 80% of utterances are kept for training, 10% for validation data, and 10% for the test, with no overlap among them.

3.2. Input representation

In our experiments, acoustic features for the conventional AAI systems are extracted from a down-sampled waveform at 16 kHz using an analysis window of length $25ms$ with frame shift of $10ms$, yielding a frame rate to match rate of the EMA recordings. Acoustic features are calculated from 40 filters, which are linearly spaced on the Mel-scale frequency axis. Log energies in the overlapping frequency bands are called filterbank energy (FBE) features. By taking the discrete cosine transform from FBEs, MFCCs can be extracted. The first 13th cepstral features, including energy, are kept and higher cepstral features are liftered to remove the fine details of the spectral envelop.

3.3. Output representation

For the articulatory space representation, instead of using EMA measurements, tract variables (TVs) [36] are employed. TVs are relative measures and suffer less from non-uniqueness [37]. We employed nine TVs, which are obtained by geometric transformations on EMA measurements. Those TVs are Lip Aperture (LA), Lip Protrusion (LP), Jaw Angle (JA), Tongue Rear Constriction Degree (TRCD), Tongue Rear Constriction Location (TRCL). In a similar way for TB and TT we have TBCD, TBCL, TTCD and TTCL, respectively.

3.4. Neural Architectures

To better assess the proposed solution, three baseline systems are built following state-of-the-art guidelines, as shown in the top two panels in Figure 1. The first and second baseline systems employ hand-crafted features, MFCCs and FBEs. The baseline with MFCC features, **base1**, consists of two BLSTM layers with 128 cells in both forward and backward directions. The baseline with FBE features, **base2**, uses a cascade of 1D convolutional layers to extract high-level features from FBEs, and two BLSTM layers with 128 cells are used to provide dynamic information to the full connected layer to predict TVs [19]. The third baseline, **base3**, is inspired from [38], which is similar to our proposed method, but it utilizes a 1D convolutional layer to extract features over a *windowed* speech signal. In that 1D convolutional layer, 256 filters with size spanning 320 samples (20ms) are used for feature extraction; next, two BLSTM layers with 128 cells in each layer followed by a dense layer are used to predict TVs. It should be noted that a batch-normalization layer was employed after the 1D convolutional layer, following [38], to prevent vanishing gradient.

In our solution, which is showed in the bottom panel in Figure 1, the filter size of the convolutional layers can be very small due to multi-stage filtering. The first layer filter size thus spans 40 samples, which is around 2.5 milliseconds (ms): the following convolutional layer has filters with a size spanning 20 samples and with decimation through the max-pooling operator, the temporal span of second convolutional layer filters are 10ms. Filtering and decimation are carried out till features at 100Hz rate, which is equal to the TVs rate, are obtained. The time span for each of the feature vectors with rate 100 Hz is equal to $70ms$ considering all of the filtering and decimation layers. In our approach, the batch-normalization layer resulted to be useless, since there were not vanishing gradient issues. The TCN contains 64 filters with length 3 and dilation rates of power two up to 256, which is bigger than the maximum input sequence length (400 samples or 4 seconds). The TCN output are passed through a 1D convolutional layers followed by time distributed fully connected layer to predict TVs.

3.5. Performance metric

To measure the accuracy of the AAI approach, Pearson's correlation coefficient (PCC) is chosen. The PCC measures the similarity of the two trajectories, and it is a normalized score which is independent of different range of speakers' articulatory movements. The PCC measure is defined as follows:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(y(i)-\bar{y})(\hat{y}(i)-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}\left(y(i)-\bar{y}\right)^2\sum_{i=1}^{N}\left(\hat{y}(i)-\bar{\hat{y}}\right)^2}}, \quad (1)$$

where $y(i)$ and $\hat{y}(i)$ are the ground-truth and estimated EMA values of the i^{th} frame, respectively; $\bar{y}$ and $\bar{\hat{y}}$ are mean values of $y(i)$ and $\hat{y}(i)$.

4. Experimental Results

In the first set of experiments, the goal is to compare and contrast the use of 1D convolutional filters to extract features directly in the temporal domain from either a windowed speech signal, i.e., **base3**, or without the windowing operation, i.e., our solution. Next, we compare the proposed method against all the three baseline systems in different experimental scenarios in terms of matching and mismatching SR conditions. All the AAI systems are speaker independent, and are evaluated both with matched and mismatched speakers for the training and testing. In the mismatched speaker scenario, the leave-one-speaker-out cross validation (LOSO) strategy is employed to carry out the assessment.

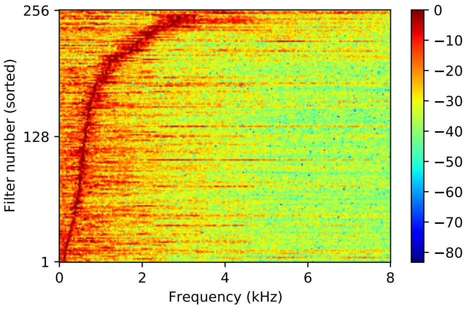

Figure 2: *The magnitude response of learned filters sorted by center frequency for* ***base3*** *system.*

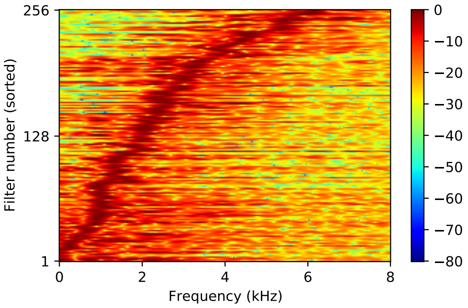

Figure 3: *The magnitude response of learned filters sorted by center frequency for the proposed method.*

4.1. 1D-CNN feature extractor

In the proposed and **base3** solutions, the first convolutional layer is extracting the features from the raw speech signal; however, a windowing pre-processing step is employed in **base3**. To better appreciate the effect of the windowing process, the characteristics of the learnt filters can be compared. To this end, the frequency response of filters is computed, and the magnitude responses are sorted by the center frequency along the frequency axis and displayed in Figure 2 for **base3**, and Figure 3 for **base3** and proposed methods, respectively. From Figure 2, it can be observed that ≈60% learnt filters' center frequency are linearly spread below 1000Hz and are non-linear above it. The highest center frequency of filters in **base3** system is less than 4000 Hz. The narrow-band magnitude response of filters can be described by the filters size which is 320 samples (20 ms). In Figure 3, due to the short filter size (2.5ms), the learnt band-pass filters have a bigger bandwidth compared to that of the **base3** system in Figure 2. Moreover, 75% of filters' center frequencies are non-linearly spread up to 3000 Hz. The center frequencies are up to 6000 Hz, which is due to short duration of the filters and therefore high frequency components of sounds do not filter-out through the filtering of first layer. The preservation of detailed information at high frequency is very useful in the estimation of TVs for high frequency sounds, such as fricatives.

Table 2: *The average PCC for different systems in the matched speaking rate condition. Spk cond indicates whether the speakers in the training and testing sets are matched or mismatched.*

Spk cond	test-SR	Proposed	base1	base2	base3
matched	N	0.84	0.83	0.80	0.81
mismatched	N	0.72	0.7	0.66	0.7
matched	F	0.79	0.79	0.73	0.78
mismatched	F	0.66	0.64	0.58	0.62
NO. Parameters		377,827	544,009	1,585,033	873,481

Table 3: *The average PCC for different systems in the mismatched speaking rate condition. Spk cond indicates whether the speakers in the training and testing sets are matched or mismatched.*

Spk cond	test-SR	Proposed	base1	base2	base3
matched	N	0.76	0.71	0.70	0.73
mismatched	N	0.65	0.52	0.56	0.61
matched	F	0.78	0.78	0.73	0.78
mismatched	F	0.68	0.67	0.64	0.66

4.2. Matched speaking rate

We now assess the effectiveness of the proposed solution in matched SR conditions. The training and test datasets have the same SR, as described Section 4, but the speaker condition, *Spk cond*, can be either matched of mismatched, as mentioned in the end of Section 4. Table 2 shows the average PCC results for different systems, where "N" and "F" stand for normal and fast SR respectively. PCC for all systems in normal SR is higher that that in fast SR. The latter is inline with what expected, since coarticulation effects are more severe in fast SR compared to those in normal SR, so capturing and tracking them is more challenging. Interesting, MFCCs allow better performance than FBEs, as observable by comparing **base1** and **base2** in Table 2. In the mismatched speaker condition, it can be observed that the system performs worse than the matched speaker condition by $\approx$0.12 in PCC for both normal and fast SR, which is expected (first and second row of Table 2). In matched speaker conditions, the proposed system attains the best results in terms of PCC and is competitive with the state-of-the-art **base1** system in fast SR. Interestingly, **base3** attains comparable or lower PCC compared to **base1** although the input features are in the temporal domain. The latter supports the discussion laid out in Section 4.1. Finally, the last row in Table 2 reports the number of parameters used in each system. A visual inspection of Table 2 allows us to argue that proposed solution attains, overall, the best results with significantly less network parameters.

4.3. Mismatched speaking rate

We now turn to the problem of performing AAI in mismatched SR. To this end, we use the systems in Section 4.2 but tested in mismatched speaking rate condition. To clear ideas: systems trained on normal SR data are evaluated on fast SR conditions, and vice versa. Table 3 summarizes the experimental evidence, in terms of average PCC, in both matched and mismatched speaker conditions. Tested on fast SR, the performance of systems trained on normal SR drops significantly compared to that obtained on normal speaking rate in Table 2. That is expected, since fast SR causes an increase in the overlap among articulators (increased coarticulation); therefore, AAI systems trained on normal SR can not model fast coarticulation movements in a proper way. However, the proposed method performance tested on fast SR achieve PCC=0.65 while the **base1** has PCC=0.52, which is a relative 20% improvement. There is no appreciable drop in PCC when systems trained on fast SR are tested on normal SR, and that is due to the fact that required information to model normal coarticulation is also available in fast SR data. By looking at the last row in Table 3, it can be observed that the results in fast SR trained model, is better when predicting the normal SR, which is another confirmation of easier prediction of TVs in normal SR which has less coarticulation.

5. Conclusion

In this work, we addressed the acoustic-to-articulatory problem is addressed in the temporal domain. Compared to conventional state-of-the-art AAI solutions based on hand-crafted short-term frequency features or windowed speech signal, 1D convolutional filters are used to extract features meaningful for the AAI task. Moreover, to match the articulatory rate, we avoid windowing, which reduce precision in capturing details at high frequency, and leverage instead decimation techniques. Moreover, a temporal convolutional network (TCN) followed by a dense layer is employed to map learned features to the TVs. Experiments are conducted on HPRC database, which provides synchronously recorded speech and EMA measurements for eight speakers. Experimental evidence demonstrates that our solution is feasible and attains top performance in mismatched speaking rate conditions, and competitive performance in matched speaking rate using however a significantly smaller amount of neural parameters.

6. Acknowledgements

This work has been supported by PRIN 2007 project nr. JNKCYZ_002.

7. 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] A. S. Shahrebabaki, N. Olfati, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Transfer Learning of Articulatory Information Through Phone Information," in *Proc. Interspeech 2020*, 2020, pp. 2877–2881. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1139

[5] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, 2013.

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

[7] 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.

[8] O. Engwall, “Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher,” *Computer Assisted Language Learning*, vol. 25, no. 1, pp. 37–64, 2012.

[9] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, “Classification of depression state based on articulatory precision.” in *INTERSPEECH*, 2013, pp. 2172–2176.

[10] S. Sahu and C. Y. Espy-Wilson, “Speech features for depression detection.” in *INTERSPEECH*, 2016, pp. 1928–1932.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] X. Xie, X. Liu, and L. Wang, “Deep neural network based acoustic-to-articulatory inversion using phone sequence information,” in *Interspeech 2016*, 2016, pp. 1497–1501.

[18] 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.

[19] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, “Sequence-to-Sequence Articulatory Inversion Through Time Convolution of Sub-Band Frequency Signals,” in *Proc. Interspeech 2020*, 2020, pp. 2882–2886. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1140

[20] R. Korin, P. Hoole, and S. King, “Announcing the electromagnetic articulography (day 1) subset of the MNGU0 articulatory corpus,” in *Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

[21] C. Qin and M. Á. Carreira-Perpiñán, “A comparison of acoustic features for articulatory inversion,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[22] P. K. Ghosh and S. Narayanan, “Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion,” *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–EL257, 2011.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1speech files available. see http://www.elsevier.nl/locate/specom1,” *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[24] C. O. Mawalim, S. Wang, and M. Unoki, “Speech information hiding by modification of lsf quantization index in celp codec,” in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 1321–1330.

[25] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A. S. Shahrebabaki, and J. Amini, “Voice conversion based on feature combination with limited training data,” *Speech Communication*, vol. 67, pp. 113–128, 2015.

[26] X. Liu, M. Sahidullah, and T. Kinnunen, “Learnable MFCCs for Speaker Verification,” in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, Daegu, South Korea, May 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03139532

[27] A. S. Shahrebabaki, J. Amini, H. Sheikhzadeh, M. Ghorbandoost, and N. Faraji, “Reduced search space frame alignment based on kullback-leibler divergence for voice conversion,” in *Advances in Nonlinear Speech Processing*, T. Drugman and T. Dutoit, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 83–88.

[28] Z.-H. Tan and I. Kraljevski, “Joint variable frame rate and length analysis for speech recognition under adverse conditions,” *Computers & Electrical Engineering*, vol. 40, no. 7, pp. 2139–2149, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790614002304

[29] X. Zeng, S. Yin, and D. Wang, “Learning speech rate in speech recognition,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform cldnns,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[31] H. Seki, K. Yamamoto, and S. Nakagawa, “A deep neural network integrated with filterbank learning for speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5480–5484.

[32] J. Fainberg, O. Klejch, E. Loweimi, P. Bell, and S. Renals, “Acoustic model adaptation from raw waveforms with sincnet,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 897–904.

[33] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks for action segmentation and detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018.

[35] 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.

[36] 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

[37] 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

[38] A. Illa and P. K. Ghosh, “Representation learning using convolution neural network for acoustic-to-articulatory inversion,” in *ICASSP*, 2019, pp. 5931–5935.