Representation Learning of Tongue Dynamics for a Silent Speech Interface

Hongcui WANG†,††, Member, Pierre ROUSSEL††, and Bruce DENBY††a, Nonmembers

SUMMARY A Silent Speech Interface (SSI) is a sensor-based, Artificial Intelligence (AI) enabled system in which articulation is performed without the use of the vocal chords, resulting in a voice interface that conserves the ambient audio environment, protects private data, and also functions in noisy environments. Though portable SSIs based on ultrasound imaging of the tongue have obtained Word Error Rates rivaling that of acoustic speech recognition, SSIs remain relegated to the laboratory due to stability issues. Indeed, reliable extraction of acoustic features from ultrasound tongue images in real-life situations has proven elusive. Recently, Representation Learning has shown considerable success in learning underlying structure in noisy, high-dimensional raw data. In its unsupervised form, Representation Learning is able to reveal structure in unlabeled data, thus greatly simplifying the data preparation task. In the present article, a 3D Convolutional Neural Network architecture is applied to unlabeled ultrasound images, and is shown to reliably predict future tongue configurations. By comparing the 3D CNN to a simple previous-frame predictor, it is possible to recognize tongue trajectories comprising transitions between regions of stability that correlate with formant trajectories in a spectrogram of the signal. Prospects for using the underlying structural representation to provide features for subsequent speech processing tasks are presented.

1. Introduction

1.1 Silent Speech Interfaces

In many real life situations – meetings, public places, etc. – it would be convenient to be able to use speech as an interface without disturbing the ambient audio environment. To this end, the Silent Speech Interface or SSI has been proposed [1], in which speech recognition and/or synthesis is performed using signals from non-acoustic sensors. Since SSIs do not require a viable acoustic speech signal, they are also of interest in noisy environments, where traditional microphones are inadequate, and in medical applications, for persons who have lost the ability to activate their vocal chords effectively due to illness or accident.

The basic outline of an SSI is shown in Fig. 1, consisting of a multi-sensor acquisition system followed by a pattern recognition engine and a means of producing speech output, either in the form of a phonetic transcription, or directly into an acoustic wave. An example of a helmet used in an SSI based on ultrasound imaging of the tongue also appears in the figure. A recent summary of the current state-of-the-art in SSI research appears in [2].

Despite the attractiveness of SSI technology, however, their deployment outside the laboratory has been hindered by the difficulty of crafting stable, viable features that enable good quality silent speech recognition. Recently, with the growth of Machine Learning, or “AI” in many fields, attention has shifted to the use of neural networks to aid in transforming SSI sensor signals into speech.

1.2 Representation Learning

The very first SSI application [3], used multilayer perceptrons to map ultrasound tongue contours to vocoder parameters. In more recent work, the use of Convolutional Neural Networks, or CNN, for SSIs has become popular [4], [5], as CNNs can replace “hand-crafted” features with automatic features obtained via convolutional kernels. Traditional AI
Representation Learning, RL, is a branch of AI in which a “representation” of underlying structure in raw, high-dimensional data is deduced [6]. Although there are many types of RL, an exciting variant is unlabeled RL, in which the expensive labeling step is unnecessary. A classic example of unlabeled RL is the Auto-encoder [7]. In an Auto-encoder – a type of Deep Neural Network – weights are trained so as to minimize the difference between raw data examples presented at the input to identical copies of these examples created at the output, see Fig. 2. In so doing, the Auto-encoder uses its hidden units to build an internal representation of structure present in the data. Once trained, the activations of the middle layer neurons may be used as features for performing, for example, recognition based on presented raw inputs. Auto-encoders have enjoyed success in a number of fields. For SSIs, auto-encoder produced features have already given good results for speech recognition in an SSI application [2].

In the present work, a different aspect of RL is exploited: the ability to learn structure from data by predicting the future, for example, predicting future frames in a video sequence, as proposed in [8]. Here, a 3D Convolutional Neural Network, or 3DCNN is used to predict future frames in ultrasound tongue videos. Similarly to the Auto-encoder example, learned structural data will become embedded in the activations of a central network layer, which can then be exploited as features in a recognition step. This article details the first step in this procedure: demonstrating that an RL architecture can indeed accurately predict future ultrasound tongue images, and is believed to be the first work linking structures discovered by an unlabeled RL architecture to observable acoustic events in SSI input data. The extraction of the features created by the 3DCNN, and their ability to improve the ultrasound SSI Word Error Rate from 14.5% (benchmark using Discrete Cosine Transform features) to 10.4% (3DCNN features), is demonstrated in a recently published companion work [9].

The datasets used are described in the next section, and the Machine Learning architecture detailed in Sect. 3. Results are presented in Sect. 4, and discussion and conclusions in Sect. 5.

2. Datasets Used

The data analyzed is from the Silent Speech Challenge (SSC) archive, consisting of over 700,000 TIMIT training images (47 lists of 50 sentences) and 35,000 WSJ0 test images (100 sentences) recorded in our laboratory in 2012 [10], in the non-verbalized punctuation manner. The speech was recorded silently, i.e., without any vocalization; there is therefore no audio track. The archive contains visual lip and ultrasound tongue images; only the tongue data were used in the current tests. Ultrasound data consists of 320x240 pixel sagittal tongue images, acquired using a 128 element 4 to 8 MHz micro-convex probe attached to a helmet like the one in Fig. 1, at a frame rate of 60 images (“frames”) per second (f.p.s.). As the probe is micro-convex (radius 12 mm), it can be placed beneath the lower jaw, quite close to the neck, and interferes little with the motion of the lower jaw. Images were reduced to 96x96 pixels before being presented to the predictors. Examples of tongue images before re-sizing are shown in Fig. 3. Tongue tip is at upper right. The hyoid bone shadow appears on the left side of the image.

3. Machine Learning Procedure

A number of architectures, including DNN and CNN, were tried before choosing a 3DCNN. Simpler architectures produced predictors without sufficient detail to detect motion. In a 3DCNN, using time as a third dimension, the input space, as well as the convolutional kernels and feature maps, are interpreted as 3-dimensional. Here, the network accepts a set of consecutive tongue images as input and attempts to predict the next image in the series, which is also the target.

The final feature map multiplicities of the chosen architecture were 1-16-32-64-32-16-1 as shown in Fig. 4. The first three layers perform 3D convolution, max pooling, and

![Fig. 2](image-url) Structure of an auto-encoder. The architecture consists of (usually) symmetric encoder and decoder sections. Examples X enter at left, and have as targets the same examples X', except in cases where noise has been added to the input [7]. Intermediate hidden layers converge to a middle or “code” layer, z, whose activations, after auto-encoder training, can be used as features for a subsequent classification or other type of problem. The z layer classically has fewer units than the inputs or outputs; however “overcomplete” auto-encoders [7], with higher dimensionality in the middle layer, as is the case in this article, are found also to produce useful representations (see discussion in text).

![Fig. 3](image-url) Video images of ultrasound images of the tongue for several different sounds, from an SSI application. The lip images in the archive are not treated in this article.
A Mean Squared Error (MSE) objective function was chosen. Other options such as Mean Absolute Error, Mean Squared Log Error, etc., gave very similar results. The entire SSC Archive data set, as described in the preceding section see was used for training and testing. A ReLu activation function in each convolutional step was used, and the Adam optimizer with a learning rate of 0.0001, Glorot Uniform weight initialization, and a batch size of 16 in the training.

4. Results

Predictors with 4, 8, and 16 images sequences as input were tested. The 8-image predictor gave substantially better results than the other two choices (always using the same baseline architecture). Over the full test set, the prediction MSE results were: 42.7 for 4 images; 46.2 for 16 images; and 40.4 for 8 images (see also Sect. 4.2). Indeed, at 60 f.p.s., 8 input images corresponds to roughly half the duration of a phoneme.

4.1 Visual Evaluation of Tongue Motion Prediction

The first test of the 3DCNN predictor was to determine if tongue movement is genuinely predicted, since at 60 f.p.s., inter-frame movements will be small. That this is the case can be observed in Fig. 5, which displays overlays of two successive original images (left half of Fig. 5) and of two successive predicted images (right side of Fig. 5); for downward tongue movement (upper half of Fig. 5), or upward tongue movement (lower half of Fig. 5).

The images shown are drawn from parts of the speech corpus in which the tongue is moving rapidly. The similarity of the predicted 9th image to the true 9th image is most apparent in the tongue contour on the upper part of each image, particularly in the tip region on the right side. This of course is also the most acoustically relevant region of the tongue. Other zones of similarity are also visible, for example on the lower right near the floor of the mouth. Although this region does not explicitly participate in articulation, its disposition might nevertheless be a valuable structural clue in a subsequent recognition step. In any event, as the learning procedure is meant to be unlabeled, masking certain areas of the image is not considered here. One may remark not only that the 3DCNN accurately represents the predicted structure of the tongue during these movements, but also that it does not capture the detailed speckle pattern observed in the overlay of the original images. This is reasonable because speckle is a noise process that is coherent only over a small local time window, and thus is not well predicted from a set of earlier frames. This lack of detail, of course, will be penalized in the MSE score of the predictor. At the same time, speckle, though an intrinsic part of ultrasound images, is not a relevant quantity from the standpoint of speech acoustics. Tongue structure, however, which the 3DCNN predicts well, is relevant for speech. These points will be further detailed in the next section, and addressed again in the conclusion.

4.2 Linking MSE Scores to Tongue Dynamics in Speech

The performance of the 3DCNN predictor can be gauged by measuring the MSE between the predicted 9th image and the true 9th image. To help interpret these results, the 3DCNN MSE is compared to that of simply using the 8th image as a predictor for the 9th image. We recall that the average MSE over the test set for this predictor is 40.4. A comparison of mean MSE per sentence over the test set, along with the corresponding MSE standard deviation, is shown...
in Fig. 6. The figure shows that: 1) averaged over a sentence, the 8th image is a better predictor of the 9th image than the 3DCNN; and 2) that the results are similar for the entire test set. Indeed, it is not until we examine the distribution of MSE within a sentence that the superiority of the 3DCNN becomes apparent. This is shown in Fig. 7, with the 3DCNN MSE score in red and that of the 8th image predictor in black. It is apparent that the 3DCNN has superior performance in the “peaks” of the 8th image predictor performance – corresponding to rapid tongue movement; and
inferior performance in the “valleys” of 8th image predictor MSE, when the tongue is nearly immobile. This observation, along with Fig. 5, is the principal result of the paper, i.e., that the 3DCNN architecture is indeed able to predict frame-to-frame tongue movement during speech.

It can be seen in Fig. 7 that the MSE of the 8th image predictor decreases at the beginning and end of the sentence. This is because the speaker adopted a rest position before and after speaking each sentence. The effect is visible in other sentences as well, as seen in Fig. 8, where MSE results

---

**Fig. 7** Mean-squared prediction error, MSE, for the 3DCNN (red) and the 8th image used as a predictor (black) on one test sentence versus frame number. The text transcription of the sentence appears at top and bottom of the picture for reference. The performance of the 3DCNN predictor has better performance when 8th image predictor MSE peaks, corresponding to vowel transitions and rapid tongue movement. Note particularly the superiority for the “e” in “penalty” and the diphthongs in “for”, “fail”, and “real”; see also the four dotted ellipses in the spectrogram of Fig. 10. Note the predictions begin before and end after the spoken text, because the tongue assumed a rest position before and after each silent speech passage.

**Fig. 8** Mean-squared prediction error MSE versus frame number for 4 test sentences, comparing the $N$th image used as a predictor (black) to a $N = 8$ image 3DCNN predictor (red), $N = 4$ image 3DCNN predictor (light green), and $N = 16$ image 3DCNN predictor (blue). In all sentences, the MSE of the $N$th image predictor (black curve) grows smaller as the tongue returns to the rest position.
from 4, 8, and 16 image predictors are also compared. Removing these 100–200 milliseconds non-speech segments from the data did not improve performance on speech segments; consequently they were retained in our analysis.

The poorer performance on “fixed” tongue images in Fig. 7 corresponds to the observation made in the preceding section, that the 3DCNN, although able to accurately predict the dynamics of the tongue, is unable to reproduce the exact speckle pattern in the original images. The speckle pattern of the 8th image, however, because it immediately precedes the 9th image temporally, is a reasonable predictor of speckle in the 9th image. Now, despite the fact that the 3DCNN may be expected to lose information during the pooling operation, it may nonetheless seem odd that the 3DCNN cannot reproduce information that it also “has”, i.e., the speckle pattern in the 8th image, which by coherence should be close to that in the predicted 9th image.

To further check this point, a non-predictive “strawman” architecture was also tested, having identical structure to the 3DCNN except that images 1–8 were used to predict image 8, not 9; that is, the image to be predicted was actually included in the input. The results of this test are presented in Fig. 9, where, to synchronize the two 3DCNNs to the same predicted image, the straw-man is understood to use images 2–9 to predict image 9. The figure shows MSE$_{3DCNN}$-MSE$_8$ versus MSE$_8$, for both the true 3DCNN and the straw-man 3DCNN, where MSE$_8$ refers to the MSE using the 8th image as a predictor. In both graphs, the left-hand part corresponds to low MSE$_8$ and thus slow tongue velocities (the tongue has moved little from the 8th to the 9th image, yet their speckle patterns remain similar), and the right-hand part, to a faster moving tongue. The figure shows that for the straw-man, although MSE is somewhat improved by including the image to be predicted in the input, the overall performance has the same behavior as the 3DCNN, presumably due to the pooling operation used in the architecture. One may also remark that the straw-man 3DCNN is in fact a type of over-complete auto-encoder. Such architectures, when trained on noisy data, are known to be unable to learn the identity mapping [7], thus explaining why the straw-man, and, apparently, the true 3DCNN as well, are unable to exactly reproduce speckle noise.

Although it has been demonstrated that the 3DCNN models tongue structure and dynamics during speech, it should be possible to attach these observations to concrete examples of acoustic speech events. This can be done by studying the relation between the events identified in the MSE plots – essentially, regions of stability and regions of high tongue velocity – and more traditional cues visible, for example, in a spectrogram of the utterance under study. The Silent Speech Challenge archive does not include an acoustic track; indeed, due to observations that articulation is different in silent speech as compared to vocalized speech [12], the Challenge data were articulated silently. Nonetheless, recordings of certain sentences from the archive, articulated in vocalized speech by the same speaker in the same manner as in [10], are available. Using the uttered text as a reference, it is possible to correlate certain MSE events in Fig. 7 with acoustic events in a spectrogram of the corresponding vocalized passage.

Speech production theory tells us that speech segments corresponding to rapid tongue movement – and hence, rapid evolution of vocal tract shape – should correspond to steepening of the slopes of the formant trajectories measured.

![Fig. 9](image-url)MSE$_{3DCNN}$-MSE$_8$ versus MSE$_8$ for the true 3DCNN (left) and the straw-man 3DCNN (right) predictors. MSE$_8$ refers to the MSE using the 8th image as a predictor. The plot on the left (true 3DCNN) is obtained directly from the curve in Fig. 7. In each plot, the left-hand side corresponds to slow tongue movement (and poorer 3DCNN performance), while the right hand side corresponds to rapid tongue movement (and good 3DCNN performance). The curves show that although the straw-man (which possesses the exact speckle information of the image it is predicting) has somewhat better MSE at low tongue velocities, as compared to the true 3DCNN (20 compared to 30 on the vertical axis), it is unable to completely exploit that information, due to loss of information during the pooling step. Other sentences show very similar behavior.
in a spectrogram [13], [14]. That this is verified in the present work is demonstrated in Fig. 10, where the sentence of Figs. 7 and 9, vocalized by the Challenge speaker, is presented as a spectrogram. As expected, the regions of highest slope of the formant F2 in Fig. 10 (red curve labeled F2 to the right of the curve), correspond well – using the four dotted ellipses as a reference – with the regions of rapid tongue movement identified in Figs. 5, 7, and 9. The effect is less visible in the F3 and higher formant curves. A second example is presented in Fig. 11. Of course, changes in the lip configuration can also affect formant slope in some cases, so care is necessary in interpreting the spectrogram. The regions highlighted in Figs. 10 and 11, however, occur away from discontinuities in the spectrogram due to labial plosives or fricatives. Finally, we note that the correlation of MSE between successive ultrasound tongue images and the evolution of acoustic spectral features has also been exploited in post-synchronization of ultrasound and audio streams in an experiment where the two streams were not synchronized a priori [15], confirming the validity of our approach. We may conclude that the 3DCNN has discovered structures in the raw ultrasound tongue data corresponding to acoustic events expected by the principles of speech production.

5. Discussion and Conclusions

Silent Speech Interfaces have been under development for a number of years but remain laboratory devices due to the difficulty of obtaining stable features in non-acoustic sensor data. It is suggested that Representation Learning, which permits features to be developed automatically in unlabeled data, may be a way of improving feature extraction for SSIs. A 3DCNN operating on 8-image ultrasound tongue sequences was found able to predict tongue movement in unlabeled data. Comparison to a spectrogram of an acoustic pronunciation of the data has allowed to identify tongue movements corresponding to acoustic events expected from speech production theory.

Comparison of the predictor performance to a simple previous-image predictor has brought to light similarities between the 3DCNN and an overcomplete auto-encoder, explaining the inability of the predictor to model speckle noise. Overcomplete auto-encoders operating on noisy data are known to produce Gabor-type filters such as orientation sensitive edge detectors, that are useful for subsequent recognition tasks [7]. The 3DCNN architecture presented here has a somewhat more complex structure however, in that the
middle layer, which is the source of “useful variables” in traditional auto-encoders, consists of 12x12 kernels rather than one-dimensional activations. Results of tests to map these kernels to lower dimensional variables to be input to a speech recognition task on the same dataset, as in [2], appear in a recent companion article [9]. Finally, it is reiterated that speckle is not relevant acoustically, and that the prediction of the 3DCNN is structurally correct even for nearly still tongue images. Nevertheless, the large role played by speckle noise in MSE may be grounds for testing different types of cost functions where speckle is less emphasized. Such tests are also underway.

Acknowledgments

The research is partly supported by the Public Applied Technology Research Programs of Zhejiang Province under Grant LGF20F020008 and the National Natural Science Foundation of China under Grant 61771333.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg, “Silent speech interfaces,” Speech Communication, vol.52, no.4, pp.270–287, 2010.
[2] Y. Ji, L. Liu, H. Wang, Z. Liu, Z. Niu, and B. Denby, “Updating the silent speech challenge benchmark with deep learning,” Speech Communication, vol.98, pp.42–50, 2018.
[3] B. Denby and M. Stone, “Speech synthesis from ultrasound images of the tongue,” Proc. ICASSP, Montréal, Canada, 2004.
[4] E. Tatulli and T. Hueber, “Feature extraction using multimodal convolutional neural networks for visual speech recognition,” Proc. ICASSP, New Orleans, USA, March 2017.
[5] K. Xu, P. Roussel, T.G. Csapó, and B. Denby, “Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images,” The Journal of the Acoustical Society of America, vol.141, no.6, EL531, 2017.
[6] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives," IEEE Trans. PAMI, special issue Learning Deep Architectures, vol.35, no.8, pp.1798–1828, 2013. doi:10.1109/tpami.2013.50.
[7] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations
in a deep network with a local denoising criterion,” Journal of Machine Learning Research, vol.11, pp.3371–3408, 2010.

[8] M. Mathieu, C. Couprie, and Y. Le Cun, “Deep multi-scale video prediction beyond mean square error,” Proc. International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2016.

[9] H. Wang, P. Roussel, and B. Denby, “Improving ultrasound-based multimodal speech recognition with predictive features from representation learning,” JASA Express Letters, vol.1, no.1, 2021. https://doi.org/10.1121/10.0003062.

[10] “The Silent Speech Challenge Archive,” https://ftp.espci.fr/pub/sigma/, 2013.

[11] This is the same architecture used in an initial study of tongue video prediction: C. Wu, S. Chen, G. Sheng, P. Roussel, and B. Denby, “Predicting tongue motion in unlabeled ultrasound video using 3D convolutional neural networks,” 2018 International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5764–5768, IEEE, 2018.

[12] V.M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, A. Colazo-Simon, C. Pillot-Loiseau, P. Roussel, C. Gendrot, and S. Quattrocchi, “Silent vs vocalized articulation for a portable ultrasound-based silent speech interface,” Proc. Interspeech, Makuhari, Chiba, Japan, pp.450–453, Sept. 26–30, 2010.

[13] R. Carré, “From an acoustic tube to speech production,” Speech Communication, vol.42, no.2, pp.227–240, 2004.

[14] K.G. Munhall, D.J. Ostry, and A. Parush, “Characteristics of velocity profiles of speech movements,” Journal of Experimental Psychology: Human Perception and Performance, vol.11, no.4, pp.457–474, 1985.

[15] S. Bakst and S. Lin, “Post-collection ultrasound-audio synchronization,” The Journal of the Acoustical Society of America, vol.146, no.4, 3081, 2019.

Hongcui Wang received the B.S. and M.S. degree in Computer Science, from Shandong University in 2003 and from Institute of the Computing Technology, Chinese Academic and Science, in 2006. She received Ph.D. degree in School of Informatics, Kyoto University, Kyoto, Japan, in 2009. She worked for Tianjin University from 2010. Since 2018, she has moved to Zhejiang University of Water Resources and Electric Power. She visited Institut Langevin, ESPCI in 2019 for one year. Her current research interests are speech recognition and signal processing.

Pierre Roussel received the Ph.D. degree in physics from the Université Pierre et Marie Curie, Paris, France, in 1991. Since 1982, he has been an Associate Professor of Electronics and Automatic Control at Ecole Supérieure de Physique et de Chimie Industrielles (ESPCI) ParisTech, Paris, France. His current research interests include the application of machine learning to various fields, classification in medical and biological applications, and dynamic modeling of natural phenomena.

Bruce Denby holds a BS from Caltech, MS from Rutgers, and PhD from University of California at Santa Barbara, all in Physics. Since 1995 he has been full professor of electrical engineering in France, currently at Sorbonne Université in Paris. His research interests are in signal processing; wireless communication; and applications of machine learning.