IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2020, accepted December 16, 2020, date of publication January 11, 2021, date of current version January 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050843

Creating Song From Lip and Tongue Videos With a Convolutional Vocoder

JIANYU ZHANG[1,2], PIERRE ROUSSEL[1], AND BRUCE DENBY[1], (Senior Member, IEEE)
[1]Institut Langevin (ESPCI Paris, PSL University, CNRS, Sorbonne Université), 75005 Paris, France
[2]Center for Data Science, New York University, New York, NY 10003, USA

Corresponding author: Bruce Denby (bruce.denby@sorbonne-universite.fr)

ABSTRACT A convolutional neural network and deep autoencoder are used to predict Line Spectral Frequencies, F0, and a voiced/unvoiced flag in singing data, using as input only ultrasound images of the tongue and visual images of the lips. A novel convolutional vocoder to transform the learned parameters into an audio signal is also presented. Spectral Distortion of predicted Line Spectral Frequencies is reduced compared to that in an earlier study using handcrafted features and multilayer perceptrons on the same data set; while predicted F0 and voiced/unvoiced flag predictions are found to be highly correlated with their ground truth values. Comparison of the convolutional vocoder to standard vocoders is made. Results can be of interest in the study of singing articulation as well as for silent speech interface research. Sample predicted audio files are available online. Source code: https://github.com/TjuJianyu/SSI_DL.

INDEX TERMS Multimodal speech recognition, convolutional neural networks, ultrasound, line spectral frequencies, silent speech interfaces, vocoder, rare singing.

I. INTRODUCTION

The past several years have seen a growing interest in multimodal speech processing, for combining audio tracks with video of the speaker to enhance speech recognition in noisy environments [1], [2]; to perform lip reading [3], [4]; or in Silent Speech Interface (SSI) applications [5]–[7]. As in many fields, multimodal speech processing has taken advantage of recent AI techniques such as Deep Autoencoders (DAE), Deep Neural Network (DNN), and Convolutional Neural Networks (CNN), for example in an ultrasound based SSI, to classify phonemes, or extract spectral quantities like F0 and Line Spectral Frequencies (LSF) [8]–[11].

A frequent a goal of multimodal speech processing is audio synthesis, which may be obtained either by following multimodal speech recognition with an HMM-based or other synthesis step (e.g., [12]); or by coupling extracted acoustic parameters with a source-filter model (e.g., [11], [13], [14]). Recently, so-called ''neural'' vocoders have begun to appear as an alternative to source-filter synthesizers, sometimes involving the use of the Generative Adversarial Networks (GAN) [15]–[17] that are now widely used in generation tasks [18]. Applications of neural vocoders to multimodal speech synthesis have begun to appear [16], [19]; however, results to date, although interesting, remain preliminary. Indeed, it is difficult for today's multimodal speech processing experiments to create the large data sets of high-quality acoustic parameters necessary for training and parameter-tuning of GAN-based and other neural vocoders.

In this work, a simpler alternative, more accessible to current multimodal speech applications, combines CNN/DAE with a multimodal source-filter synthesis module. In the method (Figure 1), a CNN/DAE architecture first learns LSF, F0, and voiced/unvoiced (U/V) flag from ultrasound tongue and visual lip images, using ground truths derived from an audio track. Subsequently, a ''Convolutional Vocoder'' uses the predicted acoustic parameters to produce an acoustic signal having properties similar to the raw audio input. The method is applied to a ''rare singing'' data set, whose results can be of importance both in the study of rare singing styles, and as an exploratory study of acoustic parameter prediction for an SSI. Spectral Distortion (SD) performance of LSF prediction using the architecture is found to improve as compared to an earlier study [11] on the same dataset; and F0 prediction, using only tongue and lip images, is excellent. The performance of the convolutional vocoder compares favorably to that of a standard MELP-MLSA (Mixed Excitation Linear Prediction – Mel Log Spectrum Approximation) coder [20], [21]. Example videos of the resulting CNN synthesis and ground truth, illustrating the ultrasound tongue and visual lip images used, are provided online at [22].

A historical overview of multimodal speech synthesis is presented in the next section, followed by an outline of the datasets used in section III. The training procedure and results

The associate editor coordinating the review of this manuscript and approving it for publication was Xi Peng.

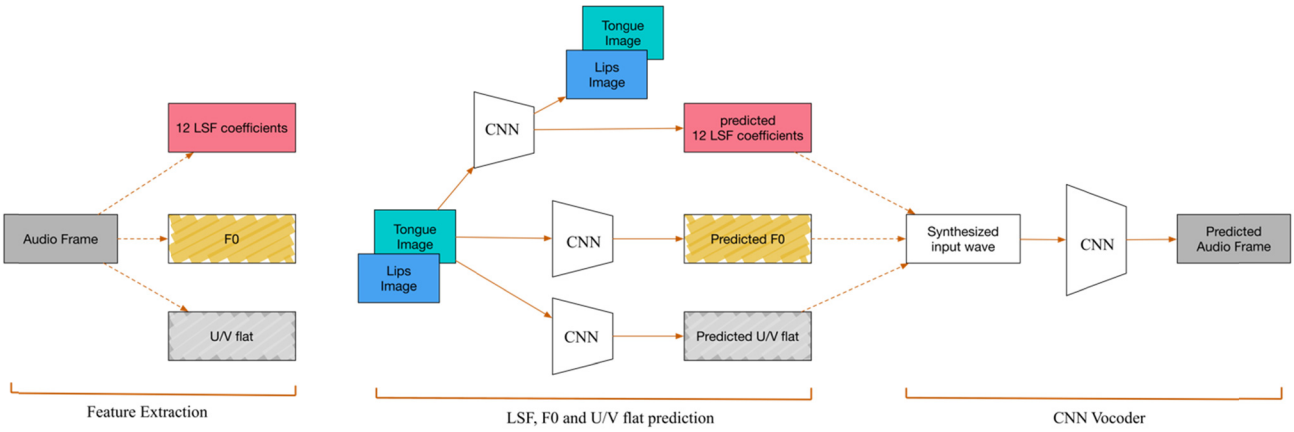


FIGURE 1. Overall pipeline. It starts from extracting LSF coefficients, F0 and U/V flag from audio frames. Then tongue and lip images are used to predict these three features separately by CNN (A dashed line indicates no back-propagation traverses it). Finally, the predicted features are used to synthesize an input wave for a CNN vocoder, which predicts audio frames.

for the CNN appear in section IV.A, and vocoder development with its results in section IV.B. Conclusions and future perspectives are discussed in the section V.

II. RELATED WORK

The interest of multimodal data for speech processing has been recognized for many years. The first mention of lip reading for speech synthesis appeared in [23] in 1985, and a review of progress in lip image based speech processing published in [24] in 2004. The first reference to ultrasound tongue imaging for speech synthesis (also in 2004 [13]) used feed-forward neural networks to map ultrasound tongue contours to acoustic parameters of a source-filter vocoder employing a white-noise activation function. In 2010 [5], a review of SSI applications exploiting a wide variety of non-acoustic sensors as multimodal inputs appeared.

More recently, an experiment [25] (2015) using tongue and lip mounted Electromagnetic Articulography coils (EMA) as input made use of an MLSA vocoder with white noise activation for synthesis. In 2017 [26], MLSA vocoder parameters were obtained from surface electromyographic (sEMG) signals, where a predicted F0 activation gave improved intelligibility compared to white noise input. In a 2019 experiment [27], ECoG brain implant signals recorded during speech were used to predict spectral quantities transformed into speech with an MLSA vocoder, producing short sentences with, in some cases, encouraging similarity to ground truth sentences.

Concerning lip reading, most recent experiments have made use of the GRID corpus [28], which constructs simply structured sentences from a limited set of words. In [29], syntheses performed using the STRAIGHT vocoder [30] with a noise activation gave about 50% word accuracy in listening tests. Increased performance on the same dataset, about 80%, was reported in [31] using CNN to produce features from lip images, and a noise-activated source/filter model. Further GRID improvements were obtained in [49] using a DAE to predict a cortically-inspired spectrogram representation of spectral parameters that preserves some F0 information, followed by direct transformation into a speech signal using an analytical technique developed in [32].

An approach at the frontier between traditional techniques and newer neural ones appears in [33], where a CNN transforms ultrasound tongue images into a spectrogram that is converted into speech using the Griffin-Lim algorithm [34]. Griffin-Lim is similar conceptually to the algorithm in [32], and is also the synthesis technique employed in the generative Tacotron vocoder [35]. In [33], word recognition rates for a set of simple commands in automated listening tests were about 60%. Finally, the generative vocoders Wavenet [15] and WaveGlow [36], have been used, respectively, in [16] (2018), to produce single-word speech outputs from ECoG brain implant waveforms; and in [19], to synthesize a set of test sentences from ultrasound images of the tongue.

In these last examples [16], [19], [33], as well as in preliminary tests of our own using Griffin-Lim [37], the spectrograms predicted from sensor data, although globally correct, lack detailed harmonic structure, giving rise to speech that, while interesting, has only moderate intelligibility. Apparently, the new vocoders, however powerful, cannot compensate for shortcomings encountered in the acoustic parameter prediction phase. This observation is in accord with the view [38] that neural vocoders may require detailed input parameters (aperiodicity, for example) not accessible in some applications; and that in [39], on singing voice synthesis, that accurately predicting acoustic feature sequences remains a key issue in vocoding. For these reasons, as well as the requirement of very large training sets for neural vocoders, we propose, in this work, an alternative approach that affords both a simpler implementation and a more transparent view of the effects of imperfect prediction of spectral and acoustic vocoder parameters.

III. DATASETS AND PRE-PROCESSING

The dataset used consisted of 5 traditional Latin and Corsican songs of 2 to 5 minutes each, for a total duration of 19 minutes [40]. The data were acquired as part of the i-Treasures project [41] that proposed new technologies for conserving

intangible cultural heritage, such as rare singing styles. Singers were instrumented with a special helmet including, besides a standard microphone: an ultrasound probe beneath the chin to record tongue movement; a camera before the mouth to capture lip movement; an electroglottograph at the neck sensitive to glottal activity; accelerometers at the bridge of the nose to measure nasality; as well as a special belt on the torso to log the respiration rate. In addition to archiving this rare cultural heritage, the data recorded by i-Treasures sensors are useful for creating models of articulation in different singing styles [42], and have also served as input to the i-Treasures Text-to-Song synthesis platform [41]. The singing data are furthermore interesting for the development speech synthesis systems for SSIs, as the longer duration of phonetic events typical in singing data can be a useful stepping stone towards more effective SSI systems.

Data were logged using a real time data acquisition system that recorded ultrasound tongue and visual lip images at 60 frames per second, as well as audio at 44.1 kHz (the other sensors mentioned earlier are not used in this study). After cleaning, 68,146 lip/tongue images and 50,087,310 audio samples were retained. A set of 5000 images was set aside for testing – identical to that used in [11] – while the rest of the data was used for training and validation. Regions of Interest (ROI) defined in the tongue and lip images were resized to 48×48 pixels. A pre-emphasis filter (a=0.95) was applied on the original audio as in [11]. Sound was downsampled to 16 kHz for the experiments; however, for the SD comparison to [11], 11.025 kHz was used, as in that work. LSF values were extracted using standard techniques and were verified to be nearly identical to those in [11]. F0 was defined as the frequency of the sine wave giving the best alignment with the audio in each frame. U/V was set to zero for frames in which the alignment procedure failed.

IV. MACHINE LEARNING ARCHITECTURES DEVELOPED, WITH RESULTS

A. CNN FOR LSF, F0, AND U/V PREDICTIONS

1) CNN TRAINING

The training architecture is illustrated in Figure 2. Resized lip and tongue ROIs feed a convolutional DAE whose central bottleneck layer provides the features used to predict 12 LSF values in parallel after a dropout layer and a dense layer. In this work, the unsupervised convolutional DAE is actually a representation learning. Due to the difficulty of obtaining of large numbers of lip and tongue images, as is often the case, unsupervised representation learning of images can help supervised prediction, especially when there are shared underlying causal factors between image reconstruction and LSF prediction [43]. For instance, both unsupervised DAE and supervised LSF coefficients prediction strongly rely on the shape of the lips and tongue, a high-level feature. The DAE learns the high-level representations from raw pixels. A classification approach in which the LSF values are first binned into 100 discrete levels was found to give superior performance compared to predicting LSF values directly. In the case of F0 and U/V, the Encoder and Predictor in with regression output and binary classification output, respectively, are used for training to gain better performance. Training was done using Tensorflow in Python. In Convolutional and Pooling layers, parameters are listed in the form of: channels @ kernel × kernel. By default, activation function and strides are set to RELU and 1, respectively. Denoising dropout is set to 0.2. Here we use Batchnorm and Dropout techniques together because of the limited small dataset and the difficult learning task.

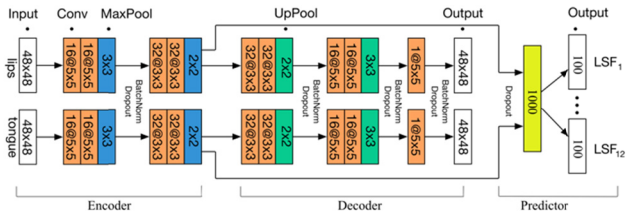


FIGURE 2. LSF training architecture. It contains: Encoder, Decoder and Predictor. The Encoder and Decoder constitute a DAE acting on resized tongue and lip ROIs, while the Predictor outputs the LSFs. For F0 and U/V training the Decoder is not implemented. U/V prediction replaces twelve 100-class classification output layers by a binary classification layer. F0 prediction further replaces the classification output layer by three regression output layers, which predict F0, Amplitude and DC offset of audio waveforms, respectively.

TABLE 1. Performance of LSF, F0 and U/V prediction.

		Frequency (kHz)	Spectral Distortion (SD) dB
LSF	CNN/DAE (this work)	16	3.80
		11.025	3.50
	DAE/MLP [11]	11.025	4.3

	Pearson Corr. Coef.	NMSE
F0	0.936	0.008

	ROC AUC	Accuracy
U/V	0.969	0.930

In train mode, we use Adam optimizer [44] with learning rate 1e-4, Glorot Uniform [45] weights initialization, batch size 512 for LSF, F0 and U/V predictions. LSF prediction uses weighted sum of MSE loss on reconstructed lips and tongues (the weights is 1e-2 for both lips and tongues) and cross entropy loss on discrete LSF levels (the weights is 1 for each LSF coefficient). Then it is trained with 50 epochs. F0 and U/V predictions use MSE loss and cross entropy loss, respectively. Both F0 and U/V are trained with early stopping on a 5% validation dataset to avoid overfitting. More details can be found at: https://github.com/TjuJianyu/SSI_DL.

2) CNN RESULTS

The performance results for prediction of the LSF, F0, and U/V flag are given in Table 1.

Performance of the LSF prediction is measured in dB of Spectral Distortion, defined as the root mean square of the

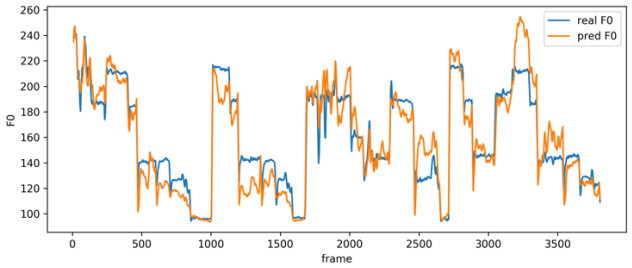


FIGURE 3. Real and predicted F0 versus frame number. Unvoiced frames, where F0 is not defined, do not appear in the plot.

differences in dB, at a fixed set of frequencies, between the LPC polynomials derived from the original and the learned LSF values [14]. In the LSF entry of Table 1, a comparison is made between the present work and [11], where DAE features were selected manually for saliency before being used to train a multilayer perceptron (MLP). The train and test datasets and pre-processing procedure used in [11] and this work were identical. An improvement of 0.8 dB is obtained using the CNN/DAE approach of this work.

For F0, a Pearson correlation score of 0.936 and Normalized Mean Squared Error (NMSE) of 0.008 were obtained. The NMSE is defined as the mean squared difference between the true and predicted values, divided by the means of the true values and of the predicted values. A high degree of F0 correlation was also observed in [10] (Pearson of 0.7 and NMSE of 0.5) where a DNN mapped tongue (only) images to F0. A plot of real and predicted F0 values from the test set of the present work appears in Figure 3.

The U/V binary classifier results are expressed in terms of ROC AUC and Accuracy. ROC AUC ranges from 0.5 to 1, with 0.5 corresponding to random guessing. An ROC AUC of 0.969 is interpreted as a 96.9% probability that a randomly chosen positive example is ranked higher than a randomly chosen negative one; while 0.930 Accuracy means an 93.0% probability of giving the correct classification result.

B. CONVOLUTIONAL VOCODER

1) VOCODER TRAINING

For multimodal synthesis applications, predicted acoustic parameters can be used to produce an audio output signal. Although LSF/LPC based vocoding works well when ground truth residuals are used as input to the synthesis step, obtaining realistic sounding vocalizations using artificial activation functions is more problematical. In [11], numerous experiments based on real and synthetic electroglottograph signals (EGG) were carried out in order to obtain reasonable sounding reproductions of the original singing audio. In this work, a CNN-based approach was used to learn to produce an output resembling the original audio. This ''convolutional vocoder'' works in the following way.

First, a ''bare'' audio signal is produced from the original, measured LSF, F0, signal phase, and U/V values. In each frame, this signal consists of a known-phase sine wave at F0, for voiced frames, or a flat waveform for unvoiced ones,

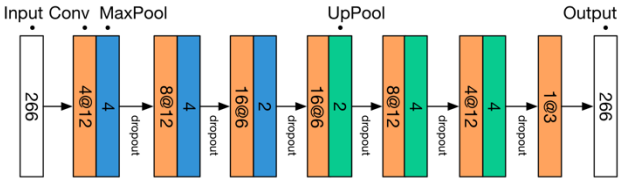


FIGURE 4. CNN vocoder architecture. Input and output are ''raw'' and original audio waveform, respectively, from one frame. Strides of MaxPool is set to 2, dropout to 0.1.

to which white noise is added before filtering by the LSF filter. A 1-dimensional CNN is then trained to ''dress'' the bare signal until it resembles as much as possible the original audio of the corresponding frame. A CNN autoencoder with a local spatial property is commonly used in audio-related tasks, such as speech enhancement [46]. In this work, we use a CNN autoencoder (Figure 4) to generate audios. For training, we use a distortion loss (MSE) as the objective function during the training of the vocoder, while other objective function, such as adversarial loss, can also be used here. An autoencoder architecture is used to synthesize audio signals from the ''bare'' audio signals. The activation functions are RELU except the last layer with a linear activation function.

In train mode, we use Adam optimizer with learning rate 1e-4, batch size 512, Glorot Uniform weights initialization, and 5000 epochs to train the neural network. More details of the neural network and the training process can be found at: https://github.com/TjuJianyu/SSI_DL.

In test mode, the *predicted* LSF, F0, and U/V flags are used to create the raw waveform. Signal phase at this point is of course unknown, but is not relevant perceptually, as long as phase continuity is assured at frame boundaries.

2) VOCODER RESULTS

This section compares the CNN vocoder to two baseline methods:

- **MELP-MLSA:** MLSA [47] is a filter based on Mel-cepstral coefficients. During the synthesis of audio, MLSA is applied to Dirac pulse trains generated at F0. Here, modified Dirac pulse activations from MELP [21] are used to improve the quality of the synthesized audio. Mel-cepstral coefficients are approximated from LPC as in [20].
- **MELP-LPC:** An LPC filter is used on the same pulse activations as in MELP-MLSA.

Another possibility would be a generative approach. The main difference between the adversarial loss in GAN and distortion loss (e.g. Mean Squared Error, or MSE) is the definition of salience. The adversarial loss connects salience with recognition ability (recognizing a true or synthesized example), while the distortion loss defines salience as total pointwise distortion between true and synthesized examples. Both adversarial loss and distortion loss can be applied based on different definitions of salience. As mentioned in section II, however, we prefer not to test generative architectures at this point. These techniques require very big datasets and significant parameter tuning to use. Based on the small

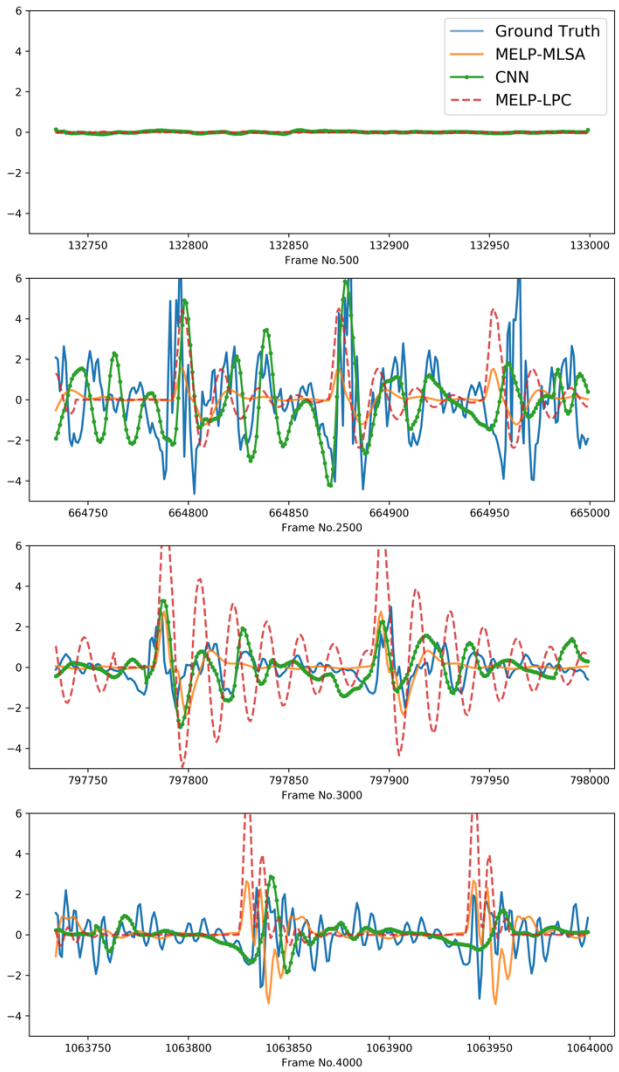


FIGURE 5. Waveforms comparison of ground truth, MELP-MLSA, CNN, MELP-LPC on frames No. 500, No. 2500, No. 3000 and No. 4000.

silent speech dataset, WaveNet for example could not be well trained and thus we do not compare with it in this project.

Figure 5 shows the comparison of CNN vocoder, MELP-MLSA, MELP-LPC and ground truth on four frames. In unvoiced frames (e.g. Frame No. 500), the methods are almost identical, while in voiced frames, the CNN vocoder waveform appears to most nearly resemble ground truth, on phones /a:/ in Frame No. 2500 and No. 3000, and /i:/ in Frame No. 4000. We note that the predicted waveforms' glottal pulses, being derived from predicted F0, should not be expected to align perfectly with ground truth.

A video of one song, showing ultrasound and lip camera modalities and using the proposed vocoder, is given in [17], and an example comparison of the CNN synthesized audio (.wav) with the baseline methods appears in [48], including a result from pure MELP [47] without LPC or MLSA filtering. It is important to remember that the predicted clips are created using *exclusively* ultrasound tongue and visual lip images as input. The behavior of the F0 prediction in Figure 3 suggests

TABLE 2. Shifted cosine similarity SCS comparison of CNN, MELP-MLSA, MELP-LPC and white noise.

	CNN	MELP-MLSA	MELP-LPC	White Noise
SCS	0.209	0.198	0.196	0.128

that we cannot, at this stage, expect a result that closely resembles the beautiful original song. Indeed a too-low F0 in certain passages makes it difficult to follow parts of the song. Errors in the U/V flag prediction as well as discontinuities in LSF values can also produce artefacts that degrade listening quality. Nonetheless it is an important first step to have produced, using the convolutional vocoder, a signal that sounds as if it might have human origin, particularly as compared with the "buzzy" quality of the MELP vocoders. Indeed the results suggest that better control of the predicted spectral parameters could lead to a much improved result. In Figure 3, for example, appropriate smoothing of F0 could produce a much less chaotic-sounding result, even if the prediction in some frames remains far from the target value. Similar procedures on the LSFs and U/V flag might also be fruitful.

We further compare the correlation between ground truth and synthesized waveforms. As discussed in section 5, the precise positions of glottal closures cannot be reproduced exactly. We thus compare the correlation between waveforms with a Shifted Cosine Similarity, SCS, method on frame i as follows:

$$SCS_i = \max_{\beta} cosine(y_{i\times f:(i+1)\times f}, \hat{y}_{i\times f+\beta:(i+1)\times f+\beta}) \quad (1)$$

$$cosine\,(a, b) = \frac{a \cdot b}{|a|\,|b|} \quad (2)$$

where f indicates frame size, $\beta(|\beta| < 0.1\text{f})$ is a shift on the synthesized audio frame $\hat{y}_{\text{i}\times\text{f:(i+1)}\times\text{f}}$, and $y_{\text{i}\times\text{f:(i+1)}\times\text{f}}$ is the ground truth audio frame. By adjusting β to maximize cosine similarity, SCS can measure the correlation between waveforms regardless of the positions of glottal closures. We constrain $|\beta|$ within a small range, 10% frame size, and present cosine similarity performances in V, which shows that the CNN method outperforms the baseline methods.

V. DISCUSSION AND CONCLUSION

A CNN combined with a DAE has been used to predict LSF, F0, and U/V flag from ultrasound tongue and visual lip images of a rare singing performance. Accuracy on LSF prediction is significantly improved compared to an existing benchmark on the same data set [11] that uses hand-engineered DAE features and MLPs. As has also been observed in other studies, [10], [13], F0 and U/V can be predicted with good accuracy from ultrasound and visual data of the vocal tract. These results can be of interest in singing articulation studies and suggest techniques that may also be applicable in the development of SSIs for ordinary speech. Finally, a convolutional vocoder trained on the original audio signal produces reasonable-sounding vocoding on test-set vocalizations [16], [48] and compares favorably to MELP benchmarks. Despite the improvements, the fidelity of

the resulting audio remains low. It is suggested that smoothing of predicted quantities could provide better performance. When parameter estimation is better controlled, extensions to some of the more recent types of vocoders [15], [36], [50] could be undertaken.

ACKNOWLEDGMENT

The authors would like to thank Aurore Jaumard-Hakoun for help in reconstructing the analysis procedure used in [11]. They would also like to thank the reviewers for providing suggestions to improve the clarity and pertinence of this article.

REFERENCES

[1] F. Tao and C. Busso, "Gating neural network for large vocabulary audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1290–1302, Jul. 2018.

[2] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 2130–2134.

[3] K. Noda, Y. Yamaguchi, K. Nakadai, G. Hiroshi Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, 2014, pp. 1149–1153.

[4] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," 2016, *arXiv:1611.05358*. [Online]. Available: http://arxiv.org/abs/1611.05358

[5] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, Apr. 2010.

[6] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2257–2271, Dec. 2017.

[7] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust articulatory speech synthesis using deep neural networks for BCI applications," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Singapore, 2014, pp. 2288–2292.

[8] 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," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. EL531–EL537, Jun. 2017.

[9] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2971–2975.

[10] T. Grosz, G. Gosztolya, L. Toth, T. G. Csapo, and A. Marko, "F0 estimation for DNN-based ultrasound silent speech interfaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 291–295.

[11] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, and B. Denby, "An articulatory-based singing voice synthesis using tongue and lips imaging," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, San Francisco, CA, USA, Sep. 2016, pp. 1467–1471.

[12] T. Hueber, E.-L. Benaroya, B. Denby, and G. Chollet, "Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Florence, Italy, 2011, pp. 593–596.

[13] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Montréal, QC, Canada, May 2004, pp. 685–688.

[14] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a silent speech interface using ultrasound imaging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toulouse, France, May 2006, pp. 365–368.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499

[16] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *J. Neural Eng.*, vol. 16, no. 3, Apr. 2019, Art. no. 036019.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[19] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis," 2020, *arXiv:2008.03152*. [Online]. Available: http://arxiv.org/abs/2008.03152

[20] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[21] K. Tokuda, T. Kobayashi, and S. Imai, "Recursion formula for calculation of mel generalized cepstrum coefficients," (in Japanese), *IEICE Trans.*, vol. J71-A, no. 1, pp. 128–131, Jan. 1988. Accessed: Jan. 15, 2021. [Online]. Available: http://www.sp.nitech.ac.jp/~tokuda/tips/mgceptr_sa2.pdf

[22] *Example Videos of CNN Synthesis and Ground Truth*. Accessed: May 2020. [Online]. Available: https://github.com/TjuJianyu/SSIWAVE/tree/master/Video

[23] N. Sugie and K. Tsunoda, "A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 7, pp. 485–490, Jul. 1985.

[24] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. Cambridge, MA, USA: MIT Press, 2004, ch. 10.

[25] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of a DNN-based articulatory synthesizer for silent speech conversion: A pilot study," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 2405–2409.

[26] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017.

[27] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[29] T. Le Cornu and B. Milner, "Reconstructing intelligible audio speech from visual speech features," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 3355–3359.

[30] 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 sounds," *Speech Commun.*, vol. 27, nos. 3–4, pp. 187–207, Apr. 1999.

[31] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5095–5099.

[32] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, Aug. 2005.

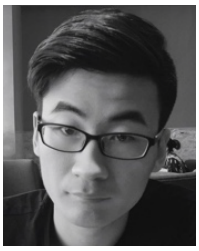
[33] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Glasgow, U.K., May 2019.

[34] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[35] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 4006–4010.

[36] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3617–3621.

[37] Y. Pang, "Développement d'algorithmes neuronaux pour synthétiser la parole á partir d'images échographiques de la langue et vidéo des lèvres," (in French), M.S. thesis, Dept. Elect. Eng., Sorbonne Université, Paris, France, 2020.
[38] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Deep neural network-based power spectrum reconstruction to improve quality of vocoded speech with limited acoustic parameters," *Acoust. Sci. Technol.*, vol. 39, no. 2, pp. 163–166, 2018.
[39] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on convolutional neural networks," 2019, *arXiv:1904.06868*. [Online]. Available: http://arxiv.org/abs/1904.06868
[40] L. Crevier-Buchman, T. Fux, A. Amelot, K. S. A. Kork, M. Adda-Decker, N. Audibert, P. Chawah, B. Denby, G. Dreyfus, A. Jaumard-Hakoun, P. Roussel, M. Stone, J. Vaissiere, K. Xu, and C. Pillot-Loiseau, "Acoustic data analysis from multi-sensor capture in rare singing: Cantu in Paghjella case study," in *Proc. 1st Workshop ICT Preservation Transmiss. Intangible Cultural Heritage, Int. Euro-Medit. Conf. Cultural Heritage (Euromed)*, Lemessos, Cyprus, 2014, p. 5ff.
[41] K. Dimitropoulos, F. Tsalakanidou, S. Nikolopoulos, I. Kompatsiaris, N. Grammalidis, S. Manitsaris, B. Denby, L. Crevier-Buchman, S. Dupont, V. Charisis, and L. Hadjileontiadis, "A multimodal approach for the safeguarding and transmission of intangible cultural heritage: The case of i-treasures," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 3–16, Nov./Dec. 2018.
[42] A. Jaumard-Hakoun, "Modélisation et synthèse de voix chantée à partir de descripteurs visuels extraits d'images échographiques et optiques des articulateurs," (in French), Ph.D. dissertation, Dept. Elect. Eng., Sorbonne Université, Paris, France, 2016.
[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[45] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[46] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.
[47] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. (Part I, Commun.)*, vol. 66, no. 2, pp. 10–18, 1983.
[48] *Example Audios of CNN Synthesis and Baseline Methods*. Accessed: May 2020. [Online]. Available: https://github.com/TjuJianyu/SSIWAVE/tree/master/Audio
[49] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2Audspec: Speech reconstruction from silent lip movements video," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2516–2520.
[50] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 839–851, 2020.



JIANYU ZHANG received the B.S. and M.S. degrees in software engineering from Tianjin University, Tianjin, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in data science with New York University, New York, NY, USA. His current research interests include machine learning, representation learning, and out-of-distribution generalization.

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

BRUCE DENBY (Senior Member, IEEE) received the B.S. degree from Caltech, the M.S. degree from Rutgers University, and the Ph.D. degree from the University of California, Santa Barbara, all in physics. Since 1995, he has been a Full Professor of Electrical Engineering with Sorbonne Université, Paris, France. His research interests include signal processing, wireless communication, and applications of machine learning.