INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An articulatory-based singing voice synthesis using tongue and lips imaging

*Aurore Jaumard-Hakoun[1,2], Kele Xu[1,2], Clémence Leboullenger[1,2], Pierre Roussel-Ragot[2], Bruce Denby[3]**

[1] Université Pierre et Marie Curie, Paris, France
[2] Institut Langevin, ESPCI ParisTech, PSL Research University, Paris, France
[3]Tianjin University, Tianjin, China
**Corresponding author*
aurore.hakoun@espci.fr, denby@ieee.org

Abstract

Ultrasound imaging of the tongue and videos of lips movements can be used to investigate specific articulation in speech or singing voice. In this study, tongue and lips image sequences recorded during singing performance are used to predict vocal tract properties via Line Spectral Frequencies (LSF). We focused our work on traditional Corsican singing "Cantu in paghjella". A multimodal Deep Autoencoder (DAE) extracts salient descriptors directly from tongue and lips images. Afterwards, LSF values are predicted from the most relevant of these features using a multilayer perceptron. A vocal tract model is derived from the predicted LSF, while a glottal flow model is computed from a synchronized electroglottographic recording. Articulatory-based singing voice synthesis is developed using both models. The quality of the prediction and singing voice synthesis using this method outperforms the state of the art method.

Index Terms: rare singing, ultrasound imaging, vocal tract modeling, Deep Neural Networks

1. Introduction

A better understanding of the mechanisms of singing performance could lead to several applications, such as improved singer trainings or singing voice synthesis applications. Recording articulatory data such as tongue and lips movements can provide much information on how the vocal tract filters the airflow during a singing performance. One non-invasive and real-time solution is to record mid-sagittal images showing tongue motion with an ultrasound probe placed beneath the chin. Lips information can be obtained with a camera. In this work, we used the combination of lips movement acquired with a camera and tongue movement recorded with an ultrasound probe, both embedded on a hyper-helmet (see Figure 1), to study singing voice articulation. Synchronized audio and electroglottographic (EGG) signals are also recorded [1]. However, using these recordings to extract vocal tract configuration is not straightforward: expert knowledge is necessary to find relevant information within the data, furthermore such multimodal recordings are difficult to integrate into a common interpretation. We hypothesized that the non-linear information in lips and tongue images could be efficiently modeled with a bimodal deep autoencoder [2], designed to process two different modalities which are lips pictures and tongue ultrasound images. Deep neural networks, very popular in signal processing since 2006 [3], are used in this work to find a data-driven shared representation between lips and tongue that contains articulatory information in an unsupervised fashion. This model is able to capture articulatory information which we attempt to map onto LSF. We used the shared representation given by the autoencoder as an input for a multilayer perceptron whose aim is to compute LSF from these given features. We hypothesized that EGG signals could be used with these predicted LSF for singing voice resynthesis. Our research is based on a Corsican singing corpus, recorded at a high frame rate (60 frames per second). Section 2 deals with datasets and data preprocessing, and a description of our deep autoencoder is given section 3. A quality evaluation of reconstructed LSF and singing voice is given section 4, compared with a linear model based on EigenLips and EigenTongues [4]. Results are discussed section 5.

2. Input data

2.1. Data acquisition

Data were acquired with a multi-sensor helmet (see Figure 1). Among many other sensors, an ultrasound probe attached on the helmet enabled the acquisition of ultrasound tongue data at 60 Hz. We used a lightweight and portable ultrasound machine, Terason T3000. Data were exported to a portable PC via Firewire (see [1] for details). In the meantime, lips data were recorded thanks to a camera at the same frame rate. Both audio and EGG were sampled at 44.1 kHz.

We investigated traditional Corsican singing synthesis. This type of singing is described in [5]. Our dataset consists of 5 traditional songs in both Latin and Corsican, including some repetitions. This base is made of 43,413 images for each modality, totalizing 723.55 seconds (about 12 minutes). This dataset includes both voiced and unvoiced sounds. Silences were not removed from the dataset.

The dataset was partitioned into a training and validation set made of 35000 images and a test set of 5000 images for each modality. The training and validation set was used as a reference for the feature extraction step: it was used to train the autoencoder, and to find the EigenTongues and EigenLips. Those models were subsequently applied on the independent test set without retraining. For the LSF prediction, the features extracted from the training and validation set were subdivided randomly to feed a multilayer perceptron (MLP), 60% of the examples were used for training and 40% for validation. The

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-385

MLP was subsequently fed using the features extracted from the independent test set, without retraining.

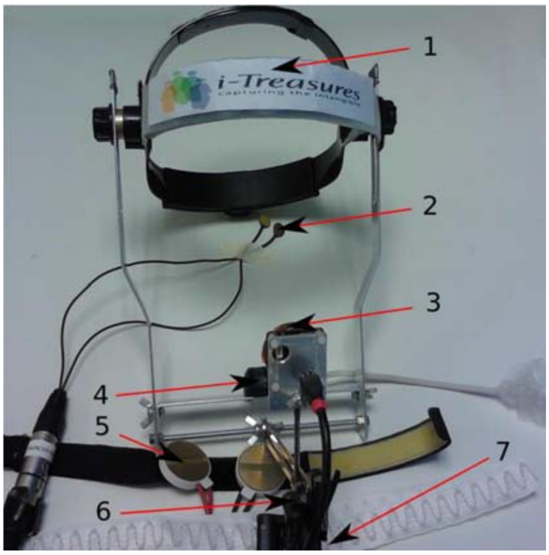

Figure 1: *Picture of the acquisition systems, which embeds several sensors. 1: multi-sensor helmet, 2: Nose-mounted accelerometer, 3: Camera, 4: Ultrasound probe, 5: Electroglottograph, 6: Microphone, 7: Breathing belt.*

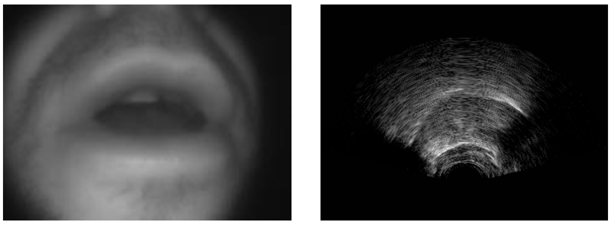

Figure 2: *Examples of tongue and lips data.*

2.2. Data pre-processing

Limiting the size of the inputs is necessary in order to keep the subsequent analyses computationally tractable. Both grayscales lips and tongue images were reduced (30x33 pixels) before being fed as an input to a deep autoencoder. Synchronized audio signals were used to compute LSF coefficients and LPC residual, and filtered with a pre-emphasis filter (α=0.95). Since 60 ultrasound frames are acquired per second, target LSF were computed every 16.7 ms: Hamming windows were applied to audio signal epochs of 33 ms each with a 50 percent overlap between epochs.

3. Automatic feature extraction

3.1. Specific constraints

We modeled two different kind of data, namely lips and tongue ultrasound pictures, using a multimodal autoencoder (see Figure 3). Previous investigations reported successful applications of multimodal autoencoders to bimodal data, for instance for audio and video data integration [2]. To ensure that the autoencoder captures the relationship between lips and tongue, we did not directly feed the encoder with a concatenation of lips and tongue data. The first step was to train separate restricted Boltzmann machines (RBMs) for lips and tongue data. Then the output of these RBMs were used as an input for another RBM, whose role was to capture a shared representation. Since we seek to reduce the size of our hidden representation as compared to the input, we added another hidden layer.

Whereas tongue images have a high signal to noise ratio, ultrasound images may be corrupted by strong speckle noise. We decided to use a denoising autoencoder [6] (see Figure 4) for ultrasound data in order improve the network robustness to speckle noise. Given an input vector $\mathbf{x}$, we created a corrupted version of $\mathbf{x}$, $\hat{\mathbf{x}}$, by adding some speckle noise (see Figure 5) to $\mathbf{x}$. During training stage, we use $\hat{\mathbf{x}}$ instead of $\mathbf{x}$ as an input of the first RBM. The output of this first RBM, $\mathbf{z}$, must be similar to $\mathbf{x}$, which means that we removed the additional noise. For this purpose, we use target $\mathbf{x}$ in the cost function $L_H(\mathbf{x}, \mathbf{z})$. Corrupted data is used only in the training stage.

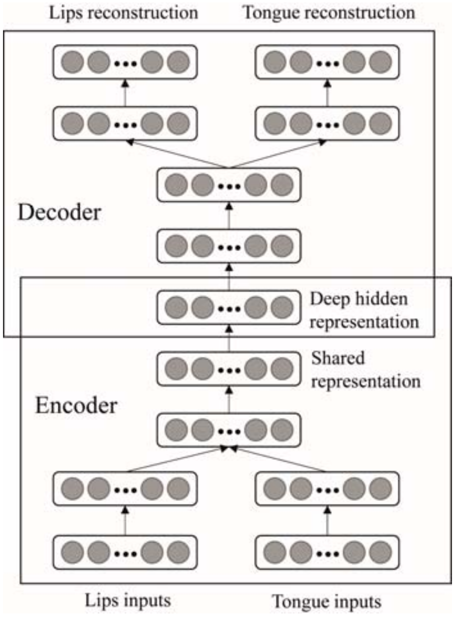

Figure 3: *structure of a multimodal autoencoder that can capture the deep hidden representation between lips and tongue data.*

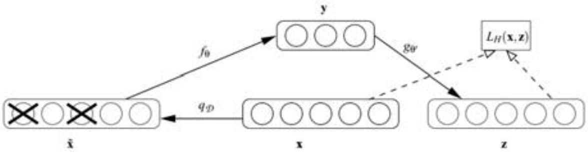

Figure 4: *Principe of denoising autoencoders. A corrupted version* $\hat{\mathbf{x}}$ *of an input vector* $\mathbf{x}$ *is used to train the first RBM and make it robust to speckle noise, from* [6].

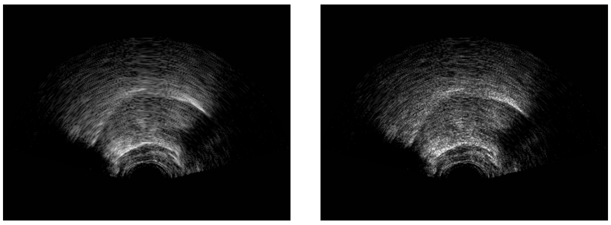

Figure 5: *Example of ultrasound image and the same image corrupted by additional speckle noise of variance 0.04.*

3.2. Autoencoder parameters, initializations and training

A deep autoencoder, built by stacking RBMs, was used to extract salient descriptors of articulatory data from lips and tongue images. We used a symmetric autoencoder with 4 hidden layers of size 200 for the separate RBMs and 100, 200 and 100 respectively for the shared RBMs. Autoencoder is first randomly initialized before RBM training. In order to improve the reconstruction performances of the autoencoder, the maximum number of epochs during training was set to 500. The size of mini-batches was set to 100. The choice of these hyperparemeters was motivated by series of test showing the relationship between parameters and quality of reconstructed LSF. An example of hyperparameter optimization is given Table 1. The number of hidden units for RBM 1 is chosen in order to minimize the mean Spectral Distorsion (see section 5.1).

Table 1 *Optimization of the number of hidden units for RBM 1 according to spectral distorsion scores.*

Number of hidden units for RBM 1	Mean Spectral Distorsion (dB)
100	4.6
200	4.3
400	4.5

4. LSF mapping

4.1. Variable selection

The mapping between hidden representation computed from lips and tongue images and target LSF coefficients was performed with one feedforward MLP network for each of the 12 LSF coefficients. Before using the hidden representation captured by the autoencoder as an input for a feedforward neural network, we performed a feature selection. The aim is to reduce the complexity of the neural network in charge of the regression of LSF coefficients. Indeed, the number of input features conditions the number of parameters of the model, and has a direct impact on overfitting risks. Feature selection was performed using the Orthogonal Forward Regression (OFR) method [7]. Input variables are ranked according to their linear correlation with a given output. Once a variable is selected, the other variables and the output are projected in a hyperplane which is orthogonal to the selected variable. This prevents the algorithm from selecting features containing the same information, using the Gram-Schmidt orthogonalization procedure. The aim of this algorithm is to select features among a large set of variables that can be strongly correlated. Random variables, referred to as "probes", were added to the feature set. Variables less correlated to the output than 95% of the probes were removed. In order to reduce even more the number of parameters of the model to control overfitting risks, we kept only a third of the remaining variable, resulting in about 30 variables per LSF.

4.2. LSF estimation

Selected variables were used as inputs of one MLP per LSF. We optimized the number of hidden units according to the lowest validation error. The 12 MLP trained for the LSF mapping had 26-33 input units, 1-7 hidden units and 1 output (each corresponding to one of the 12 LSF coefficients). Because LSF coefficients are not in the range [0-1], we used a sigmoid activation function between input and hidden layer and a linear activation function between hidden layer and output.

4.3. Comparison with state of the art methods

In order to validate our nonlinear feature extraction method, we compared its performances to the performances achieved with a linear state-of-the art method based on the extraction of the EigenLips and EigenTongues [4]. Inspired by EigenFaces in [8], this method uses Principal Components Analysis performed on tongue and lips images. The idea is to create a finite set of orthogonal images, named EigenLips and EigenTongues. Every tongue or lips configuration can be represented in this space by projection over EigenTongues and EigenLips. This representation encodes significant structural informations about lips and tongue [4].

5. Quality assessments

5.1. Quality measurements

Quality measurements were estimated on the 5000 examples from the independent test set, without retraining. Output predicted LSF were most of the time in the range of original LSF. Direct comparisons between true LSF and their predicted version using either EigenLips and EigneTongues or Multimodal Autoencoder (see Figure 6 and Figure 7) is completed by a numerical evaluation, the so-called spectral distorsion SD.

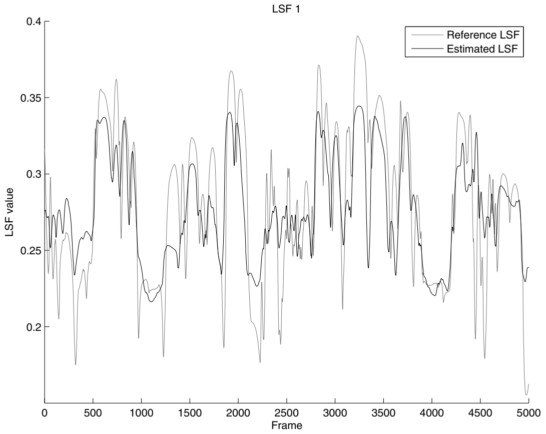

Figure 6 *Comparing true vs. estimated LSF in the test set using Multimodal Autoencoder on the first LSF.*

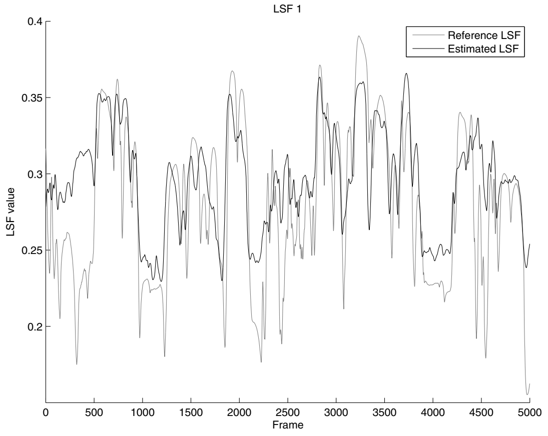

Figure 7 *Comparing true vs. estimated LSF in the test set using EigenLips and EigenTongues on the first LSF.*

Spectral distorsion is a measure in decibels of the distance between two spectra A and A'. Its expression is given below:

$$SD = \langle \sqrt{\frac{1}{(n_1 - n0)}} \sum_{k=n_0}^{n_1-1} \left(10 \log_{10} \left(\left| \frac{A\left(e^{\frac{j2\pi k}{N}}\right)}{A'\left(e^{\frac{j2\pi k}{N}}\right)} \right|^2 \right) \right)^2 \rangle \quad (1)$$

Where $n_0 = 6$ and $n_1 = 200$, which means power spectra A and A' are compared in the range 129-4307 Hz, with frequency bins of 21.5 Hz. We also use a differential MOS score, defined with a reference “transparent” distortion of 1dB in [9] by the expression:

$$MOS = 3.56 - 0.8\,SD + 0.04\,SD^2 \quad (2)$$

$$\Delta MOS(SD) = MOS(SD) - MOS(1dB) \quad (3)$$

Spectral distortion and differential MOS scores are given in Table 2.

Table 2 *Spectral distortion and differential MOS scores for our two methods and a reference distorsion of 1 dB, transparent for listeners.*

Method	Spectral distortion (dB)	Differential MOS
“Transparent”	1	0
Multimodal Autoencoder	4.3	-1.9
EigenLips and EigenTongues	5.2	-2.3

These results demonstrate that our new method outperforms state of the art methods on a dataset made of rare traditional singing. Using predicted LSF can be used to investigate the specific influence of the position of the tongue or the roundness of the lips used in this technique.

5.2. Singing voice resynthesis

Predicted LSF can be converted into LPC coefficients to build a vocal tract model. We can use the source-filter model for voice synthesis. We used different types of sources in order to compare intelligibility of the reconstructed sound. Since true LSF were computed with audio signals, we can use LPC residuals as source reference. Noise activation will produce whispered voice. EGG signals could not be used directly as a source: this correlate of the glottal activity is too indirect and noisy. We used some source information extracted from the derivative of EGG signals (dEGG) using the *Voicebox* toolbox [10] to estimate the parameters of a glottal flow model based on the CALM model presented in [11]. CALM parameters such as fundamental frequency, open quotient and asymmetry coefficient were derived from EGG signals. We compared this source with a direct application of dEGG as an activation for synthesis (which led to an unnatural voice resynthesis). Figure 8 shows the overall architecture of our synthesizer. Listening tests were conducted on a population of 84 subjects, with the aim of comparing different audio samples (see additional materials DAE.wav, Eigen.wav and Ref.wav, respectively autoencoder, EigenLips/EigenTongues and true LSF synthesized with residual source) These tests confirm that the deep autoencoder outperforms linear models, and that the glottal flow model we used, as expected, allows a much better resynthesis than dEGG or white noise (and in some cases tends to reach the quality of a reconstruction using residuals).

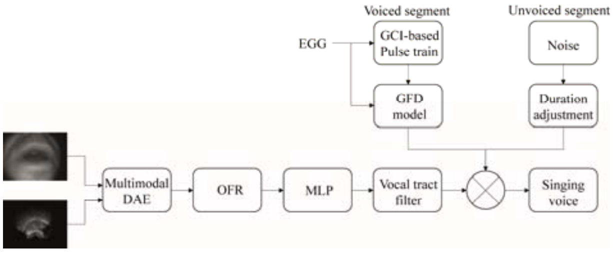

Figure 8 *Schematic view of the synthesis process. Articulatory images are used to predict LSF values using a deep autoencoder (DAE). Orthogonal Forward Regression is used to to select the best features. A Multilayer Perceptron (MLP) predict LSF values. According to a voiced/unvoiced criterion, excitation signal is either a glottal flow derivative model (GFD) or noise.*

6. Conclusion and perspectives

In this work, we presented a novel method of articulatory-based singing voice synthesis, based on the use of a multimodal autoencoder. The use of articulatory information can serve both pedagogical and artistic issues. As expected, the multimodal autoencoder is able to extract salient features from lips and tongue images, even if the formalism does not at this point allow us to understand the precise character of the extracted information. These features are used to predict LSF values using an MLP as in [9] and [4]. This study shows that our multimodal model outperforms EigenLips and EigenTongues, which is not surprising if we consider that EigenLips and EigenTongues are based on linear PCA whereas Multimodal Autoencoder is non-linear. The excitation signal is synthesized using CALM model, whose parameters are estimated from EGG signals. Synthesis using our multimodal architecture and our synthetic glottal flow models are of satisfying quality.

7. Acknowledgements

This work was funded by the European FP7 i-Treasures project (FP7-ICT-2011-9-600676-i-Treasures).

8. References

[1] S. K. Al Kork, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, L. Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone, K. Xu and B. Denby, "A Multi-Sensor Helmet to Capture Rare Singing, an Intangible Cultural Heritage Study," in *Proceedings of 10th International Seminar on Speech*, Cologne, 2014.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng, "Multimodal Deep Learning," in *ICML*, Washington, 2011.

[3] G. E. Hinton and S. Osindero, "A fast learning algorithm for deep belief nets," *Neural Computation,* vol. 18, 2006.

[4] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel-Ragot and M. Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honulu, Hawaii, USA, 2007.

[5] L. Crevier-Buchman, T. Fux, A. Amelot, S. K. Al 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 on ICT for the Preservation and Transmission of Intangible Cultural Heritage, International Euro-Mediterranean Conference on Cultural Heritage (Euromed2014)*, Lemessos, Cyprus, 2014.

[6] 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," *The Journal of Machine Learning Research,* vol. 11, pp. 3371-3408, 2010.

[7] G. Dreyfus, Neural Networks. Methodology and Applications, Berlin Heidelberg: Springer-Verlag, 2005.

[8] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[9] B. Denby, Y. Oussar, G. Dreyfus and M. Stone, "Prospects for a Silent Speech Interface Using Ultrasound Imaging," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[10] M. Brookes, «VOICEBOX: Speech Processing Toolbox for MATLAB,» Department of Electrical & Electronic Engineering, Imperial College, 1991. [En ligne]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. [Accès le 2016].

[11] B. Doval, C. D'Alessandro et N. Henrich, «The voice source as a causal/anticausal linear filter.,» chez *Voice Quality : Functions, Analysis and Synthesis VOQUAL'03*, Geneva, 2003.