Real-time Control of a DNN-based Articulatory Synthesizer for Silent Speech Conversion: a pilot study

Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, Blaise Yvert

Real-time Control of a DNN-based Articulatory Synthesizer for Silent Speech Conversion: a pilot study

Florent Bocquelet[1,2,3,4], Thomas Hueber[4], Laurent Girin[4], Christophe Savariaux[4], Blaise Yvert[1,2,3]

[1] Inserm, Clinatec, U1167, Grenoble France
[2] Univ. Grenoble Alpes, Clinatec, U1167, Grenoble, France
[3] CEA, LETI, Clinatec, Grenoble, France
[4] CNRS/Univ. Grenoble Alpes, GIPSA-Lab, Grenoble, France

florent.bocquelet@cea.fr

Abstract

This article presents a pilot study on the real-time control of an articulatory synthesizer based on deep neural network (DNN), in the context of silent speech interface. The underlying hypothesis is that a silent speaker could benefit from real-time audio feedback to regulate his/her own production. In this study, we use 3D electromagnetic-articulography (EMA) to capture speech articulation, a DNN to convert EMA to spectral trajectories in real-time, and a standard vocoder excited by white noise for audio synthesis. As shown by recent literature on silent speech, adaptation of the articulo-acoustic modeling process is needed to account for possible inconsistencies between the initial training phase and practical usage conditions. In this study, we focus on different sensor setups across sessions (for the same speaker). Model adaptation is performed by cascading another neural network to the DNN used for articulatory-to-acoustic mapping. The intelligibility of the synthetic speech signal converted in real-time is evaluated using both objective and perceptual measurements.

Index Terms: articulatory speech synthesis, deep neural networks, EMA, silent speech

1. Introduction

Building a device for oral speech communication without vocalization, i.e. a so-called 'silent speech interface' (SSI), has emerged as a new research field in the last decade [1]. SSI can be used to preserve conversation privacy and for communication in noisy environment (for which the audio signal is not exploitable). As opposed to modal speech, silent speech involves normal articulation but no vocalization. Therefore, SSI could be used after a total laryngectomy, as an alternative to esophageal and tracheoesophageal voices. Different approaches have been proposed to monitor the articulatory activity during silent speech, such as ultrasound and video imaging [2], surface electromyography (sEMG) [3], permanent-magnetic articulography (PEMA) [4], stethoscopic (NAM) microphone [5]. Several studies have addressed the problem of 'silent speech recognition', i.e. word sequence identification from silent articulation, under different modalities including ultrasound [6], sEMG [7], NAM [8], or PEMA [9]. Other studies have addressed the problem of 'silent speech conversion', i.e. direct reconstruction of a synthetic speech signal from silent articulation, without any restriction on the vocabulary (such as [10] with ultrasound and [11] with sEMG).

The present study addresses this latter problem and focuses on the 'real-time' processing of a continuous articulatory data flow. The term 'real-time' means that the delay between the articulatory movements and the synthetic speech signal has to be constant, and as short as possible. This way, the silent speaker can rely on the synthetic speech as an auditory feedback and exploit it to regulate his own production. According to the literature on delayed auditory feedback [12], the latency should be no greater than about 50 ms. A larger latency might generate a conflict between kinesthesic and auditory feedbacks. To our knowledge, such 'real-time' conversion of full sentences from silent articulation to audible speech has so far not been described in the literature.

In this study, we focus on the real-time control of a silent-to-modal speech conversion system, independently from the technology used to capture silent articulation. To that purpose, we built a system based on 3D electromagnetic-articulography (EMA), which, although invasive and not portable, is appropriate for this proof-of-concept study. EMA directly provides clean motion data on speech articulators with both good spatial and temporal resolution. In line with our previous work [13], a deep neural network (DNN) is used to convert EMA trajectories into spectral parameters. Audio synthesis is achieved using a real-time implementation of the MLSA vocoder [14], excited by white noise (hence producing whispered-like speech). As mentioned in [15] (in the context of EMG-based recognition), adaptation of the articulatory-acoustic model is necessary to account for possible inconsistencies between the initial training phase and practical usage conditions. These inconsistencies can be due to differences in the articulatory patterns of silent and modal speech [16] [17], or differences between sensor setups across sessions. In this study, we address more specifically this latter problem. We describe a supervised calibration method to adapt the articulatory-acoustic DNN to a new configuration of the EMA sensors, for the same speaker. The real-time control of the proposed SSI is evaluated for one speaker (pilot study). The quality of the reconstructed speech is assessed using both objective and perceptual listening tests.

The article is organized as follows. Section 2 describes the general characteristics of the proposed SSI system. Section 3 presents its evaluation using both objective and perceptual listening tests. Section 4 discusses these results and future work.

2. Methods

2.1. Reference session

2.1.1. Articulatory-acoustic database

In order to perform a statistical articulatory-to-acoustic mapping, articulatory data from the reference speaker were first recorded synchronously with audio signals using the NDI Wave system, during a first reference session. Seven 3D coils were glued on the tongue tip, blade, and dorsum, as well as on the upper lip, the lower lip, the jaw and the soft palate. Sequences of articulatory features were down-sampled from 400 Hz to 100 Hz, and 3D coordinates were projected in the midsagittal plane. The recorded database consisted of 712 sentences of variable length (4531 words in total). The speech signal was recorded at 22,050 Hz and parameterized by 25 mel-cepstrum coefficients using SPTK *mcep* tools (frame length 220, all-pass constant 0.455, linear interpolation was used in order to keep a sampling rate of 100Hz) [18]. In the following, all data obtained during this recording session will be referred to as the *reference data*.

2.1.2. Articulatory-to-acoustic mapping

The articulatory-to-acoustic mapping was performed by a deep neural network (DNN) trained on the reference data. We refer the reader to our previous paper for more theoretical information about this articulatory-to-acoustic mapping model [13]. Differences with our previous work lie in some parameters adjustment: in the present study, the neural network had 3 hidden layers of 200 units each, and 10 consecutive articulatory frames were concatenated in one single feature vector to take into account the dynamic properties of speech (only past frames were considered in order to not introduce any supplementary delay). Also, we used leaky rectified linear units [19] instead of more classical units like logistic of hyperbolic units, which allowed faster convergence and lower training error for the same network architecture.

2.2. Adaptation and real-time experiment

The adaptation process was divided into two parts: a calibration step during which an articulatory-to-articulatory mapping was estimated to map the EMA trajectories of the on-line session onto those of the reference off-line session, and the real-time control testing part, during which the articulatory synthesis was performed in real-time according to the subject articulatory movements in silent speech condition.

2.2.1. Articulatory-to-articulatory mapping

The calibration step is necessary in order to take into account differences across the reference session and the test session in terms of sensor positioning. In the present study, the speaker was the same in both sessions, but the number of sensors was different (from 7 to 6, no sensor was placed on the upper lip), and no particular attention was given to place each of them in the exact same position. Thus, the goal of this calibration was to create a so-called 'articulatory-to-articulatory' mapping that mapped data from the new sensor configuration to the reference one. In this calibration step, the speaker was asked to synchronously repeat a subset of 50 short sentences extracted from the reference acoustic-articulatory database, such that reference articulatory trajectories were known for these sentences. Each sentence was presented three times at a fixed pace, and the articulatory data was recorded only during the last repetition.

First, the recorded 3D coordinates were projected in the mid-sagittal plane. Then, in order to compensate system and subject latencies, a global delay between reference and new articulatory data was estimated. This estimation was done by minimizing the mean-squared error when fitting a linear model between reference and new articulatory data with different delays (preliminary experiments showed that this alignment procedure gave similar result to a DTW-based procedure). Finally, a neural network (with 2 hidden layers of 50 leaky rectified units) was trained on the calibration data in order to perform the articulatory-to-articulatory mapping.

2.2.2. Real-time control

For the real-time experiment, both the DNN used for articulatory-to-articulatory mapping and the DNN used for articulatory-to-acoustic mapping were cascaded in order to directly map new articulatory features to reference acoustic features. The mel-cepstrum coefficients obtained by the articulatory-to-acoustic mapping were then converted to audible sounds using the MLSA filter [14], with a white noise excitation signal. The articulatory data streaming, the mapping and the MLSA filter were all implemented within the Max/MSP environment (https://cycling74.com) for real-time processing. Special attention was given to audio settings in order to minimize the audio chain latency and obtain a delay inferior to 50ms. The speaker was then asked to silently articulate a set of sentences which were not part of the dataset used to train both articulatory-to-acoustic and articulatory-to-articulatory mappings. During this silent speech period, the speaker was given the synthesized auditory feedback through headphones.

3. Results

3.1. Evaluation methods

3.1.1. Objective evaluation using automatic speech recognition

Results were objectively evaluated using two different Hidden Markov Models (HMMs) for phonetic decoding: the first one was trained on articulatory data of the reference speaker (in the following, "articulatory HMM"), and the second one was trained on the acoustic data of the reference speaker (in the following, "acoustic HMM"). Both were trained from data obtained in the reference session. That way, the articulatory-to-articulatory mapping could be directly evaluated using the articulatory HMM, and the final speech synthesis result using the acoustic HMM. Both HMMs were trained using a standard procedure of context-dependent triphone tied-state HMM using the HTK toolkit [20]. The recognition accuracy (defined as $\boldsymbol{Acc\%} = 100 \cdot (N-D-S-I)/N$, where N is the total number of phones in the test set, S, D and I are the number of substitutions, deletions and insertions, respectively) was used as a measurement of the accuracy of the respective mappings. This approach was preferred to a more classic calculation of the mean-squared error (for articulatory features), or to the mel-cepstral distortion (for acoustic features) between reference and synthetic trajectories. The goal was more to evaluate the segmental intelligibility of the reconstructed

speech rather than absolute similarity with the reference speaker's voice.

For each evaluation of a data corpus, chance levels were estimated by evaluating randomly generated data with similar characteristics as the reference data (mean and standard deviation). The word insertion penalty was tuned on a validation dataset and remained fixed for all the experiments.

3.1.2. Subjective evaluation using listening tests

Ten subjects participated to an intelligibility test. All participants were French native speakers with no hearing impairment. The presented stimuli consisted of 10 French vowels /a/, /i/, /u/, /o/, /œ/, /e/, /y/, /ã/, /ɛ̃/, /ɔ̃/, and 30 vowel-consonant-vowel (VCV) pseudo words made of the 18 consonants /p/, /t/, /k/, /f/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /z/, /ʒ/, /m/, /n/, /r/, /l/, /w/, /j/, in /a/, /i/, /u/ contexts, all included in the same sentence: "Tu t'appelles *VCV*, c'est ça?" ("Your name is *VCV*, right?"), where *VCV* was replaced by each pseudo-word. Two conditions were tested: offline synthesis using reference speaker data only, and synthesis from the real-time experiment. Since the synthesized speech was completely unvoiced, a pair of consonant differing only from the voiced/unvoiced feature (but involving almost the same articulatory gestures) was pooled together, resulting in the 6 following categories: {/p/, /b/}, {/t/, /d/}, {/k/, /g/}, {/f/, /v/}, {/s/, /z/}, {/ʃ/, /ʒ/}.

In total, each participant had to identify 128 sounds played in random order at the same sound level. Participants were seated in quiet environment and instructed that they would be listening to isolated vowels or VCV sequences. For each utterance, they had to pick the corresponding vowel in the case of an isolated vowel, or the middle consonant in the case of a VCV sequence. They were told that some of the sounds were difficult to identify, and thus to choose the closest sound among the offered possibilities. No performance feedback was provided during the test. The recognition accuracy was defined as $Acc\% = R/N$ with R the number of correct answers for the N presented sounds of the test.

3.2. Adapted articulatory-to-acoustic mapping

The quality of the 'adapted' articulatory-to-acoustic mapping was evaluated using the acoustic HMM, both on the calibration corpus and real-time corpus, and using a listening test for the real-time corpus only ("Synth." line in Table 1 and Table 2). In order to obtain results comparable to the listening test results, grammar constrains were imposed to the HMM recognizer when evaluating the real-time corpus, so that it could only recognize an isolated vowel or a sentence of the type "Tu t'appelles *VCV*, c'est ça?", where *V* could be either /a/, /i/, or /u/, and *C* any consonant, then keeping only the recognized vowel or consonant as final result of the recognition. In that case, chance level was estimated by assigning random classes to the items.

Table 1 summarizes the objective evaluation results on the calibration corpus. The phones were correctly recognized with an accuracy of 62.42%, well above the chance level ($p < 10^{-6}$; Fischer's exact test) and below the accuracy obtained on the reference data ($p < 10^{-6}$).

On the real-time corpus, 61.5% of the phones were correctly recognized by the acoustic HMM ("Synth" line in Table 2), which was well above the chance level ($p < 10^{-6}$) and below the accuracy obtained on the reference data ($p = 4.10^{-4}$). Finally, subjects could correctly identify 78.8% of the phones, which was well above chance level ($p < 10^{-6}$).

The confusion matrices of the perceptual listening test (Figure 1 and Figure 2) reflect the global good accuracy of the synthesizer: 8 out of 10 vowels have recognition rate superior to 70% (vowels /a/, /y/ and /o/ were systematically identified correctly). Eight out of 12 consonants achieved 70% with a minimum of 60% (consonant /r/ was systematically identified correctly). Main confusions for vowels concern nasal vowels (that do not exist in English), which are /ã/, /ɛ̃/ and /ɔ̃/. Thus, /ã/ was often confused with /ɔ̃/ (50%), /ɛ̃/ with /a/ (37.5%) and /ã/ with /o/ (30%), whereas there was almost no confusion between the nasal vowels and the corresponding non-nasal vowels (respectively /a/, /e/, and /o/), since velum position was recorded. For consonants, minor confusions are made between consonants that differ by the nasality feature (such as {/p/, /b/} vs. /m/, and {/t/, /d/} vs. /n/), which were the consonants with lower recognition rate (<70%). Other confusions remain difficult to interpret.

	Articulatory HMM	*Acoustic HMM*
Ref.	98.01 %	97.68 %
Synth.		62.42 %
Exp.	59.93 %	50.33 %
Chance	23.6±0.4 %	16.9±0.3 %

Table 1. *Automatic recognition accuracy on the calibration corpus (Ref. corresponds to reference data, Synth. to articulatory-to-acoustic mapping using reference articulatory data, and Exp. to real-time experiment data)*

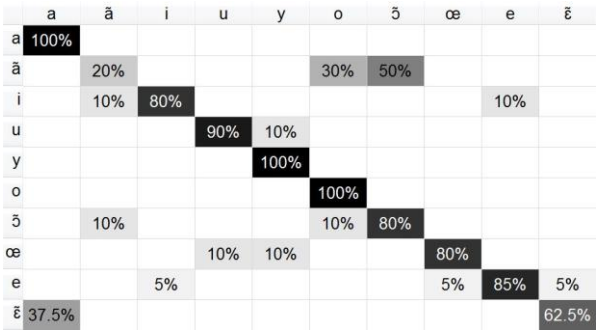

Figure 1: *Vowels confusion matrix for the articulatory-to-acoustic mapping (listening test).*

Figure 2: *Consonants confusion matrix for the articulatory-to-acoustic (listening test).*

3.3. Articulatory-to-articulatory mapping

The quality of the articulatory-to-articulatory mapping was evaluated on the calibration corpus ("Exp." line in Table 1). The articulatory HMM evaluated the output of the articulatory-to-articulatory mapping when applied on the calibration data. The output of this first mapping was then fed to the articulatory-to-acoustic mapping, the output of which was evaluated by the acoustic HMM.

On the articulatory data, the phones were correctly recognized with an accuracy of 59.93%, which was well above chance level ($p < 10^{-6}$) and below reference data score ($p < 10^{-6}$). On the acoustic data, the accuracy was 50.33%, which was again well above chance level ($p < 10^{-6}$) and below both the reference data score ($p < 10^{-6}$) and the articulatory-to-acoustic mapping score ($p < 10^{-6}$).

3.4. Real-time control

Finally, the results of the real-time experiment were evaluated in the same way as the articulatory-to-articulatory mapping using HMM recognizers (with the same grammar constrains), and a listening test ("Exp." line in Table 2).

The articulatory HMM achieved 64.6% of recognition rate, which was well above chance level ($p < 10^{-6}$) and not significantly different from the reference data score ($p > 0.3$). The acoustic HMM achieved 49.2%, which was well above chance level ($p < 10^{-6}$) and below reference data results ($p < 10^{-5}$), while not significantly different from the articulatory-to-acoustic score ($p > 0.2$). The subjects correctly identified 56.8% of the phones, which was significantly different from chance ($p < 10^{-6}$) and below the articulatory-to-acoustic results ($p < 10^{-6}$). It is interesting to note that this latter result was not consistent with the acoustic HMM evaluation, which suggests that even if HMM accuracy is a good measure of intelligibility, it is still not as good as a listening test by human subjects.

The confusion matrices of the listening test (Figure 3 and Figure 4) reflect the differences with the articulatory-to-acoustic results: 4 out of 10 vowels, and 4 out of 12 consonants, achieved more than 70% of accuracy. Vowels /a/ and /i/, and consonant {/f/, /v/} were systematically identified correctly by all participants. For vowels, confusions between nasal and non-nasal vowels is increased (80% of /ã/ were recognized as /a/, 37.5% of /ɛ̃/ as /œ/ and 25% as /e/ and as /i/, and 70% of /ɔ̃/ as /ã/). /ɛ̃/ was never identified correctly, as well as /y/ that was systematically recognized as /i/. This could be due to different articulatory patterns between modal and silent speech (since the calibration was made on modal speech), or more likely to poor mapping of lip features due to the absence of the upper lip sensor in the test session, since /i/ and /y/ mostly differ by lips protrusion. Vowel /e/ was often confused with /i/ (40%) that differ mostly by lips aperture, suggesting again a poor mapping of lip features. Substitutions of {/k/, /g/} by /r/ (35%) could be explained by similar tongue shapes in the mid-sagittal plane. Many confusions were made between alveolar sounds {/t/, /d/} and {/s/, /z/} (40%). Some confusions were made between fricative sounds {/s/, /z/} and {/f/, /v/} (25%). Other confusions remain difficult to interpret and may be due to articulatory-to-articulatory mapping errors, like {/p/, /b/} or /m/ confused with {/f/, /v/} (31.7% and 23.3% respectively).

	Articulatory HMM	*Acoustic HMM*	*Listening Test*
Ref.	73.8 %	89.2 %	
Synth.		61.5 %	78.8±5.4%
Exp.	64.6 %	49.2 %	56.8±8.7%
Chance	7.4±2.9%		

Table 2. *Percentage of correct results on the real-time corpus, for the 3 different evaluation methods*

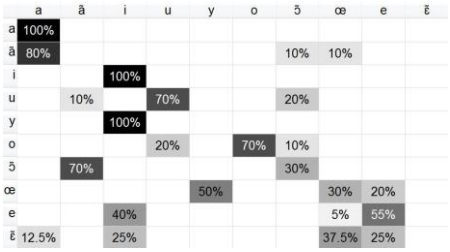

Figure 3: *Vowels confusion matrix of the real-time experiment (listening test).*

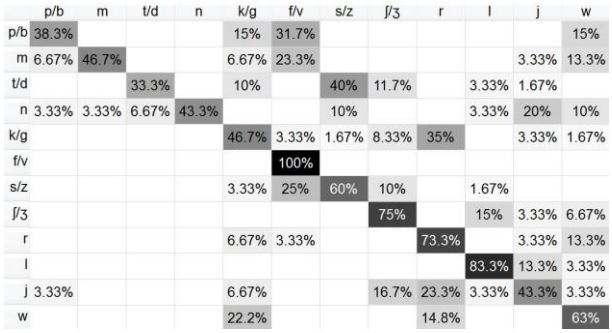

Figure 4: *Consonants confusion matrix of the real-time experiment (listening test).*

4. Conclusion

In this paper we present an approach to control an articulatory synthesizer in real-time for silent speech conversion. We propose an adaptation method based on neural networks to take into account differences of sensor number and positions across different sessions. This method was then applied in real-time with a speaker articulating sentences while being given the articulatory synthesis feedback through headphones. Results of the real-time experiment were compared to the results obtained with reference articulatory data. Objective and subjective evaluations provided phone recognition accuracy far above chance level, reaching about 57% in a perceptual listening test. However, comparison with the score obtained using reference articulatory data (about 79%) pointed out that significant part of intelligibility was lost. Confusion matrices suggest that this is mostly due to articulatory-to-articulatory mapping errors, mostly on lip features likely due to the absence of the upper lip sensor in the test session. This mapping mismatch might also be worsen by the fact that the articulatory-to-articulatory and articulatory-to-acoustic mapping were cascaded but not adapted to the new data as a whole. The fact that the articulatory HMM systematically achieved better recognition rate than the acoustic HMM both on calibration and real-time data, while no significant differences were observed on reference data, suggests that some intelligibility loss was directly caused by the synthesizer. Using an HMM-based articulatory synthesizer [10] in future studies may lead to better synthesis results.

5. References

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

[2] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, and M. Stone, "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips," *Speech Commun.*, vol. 52, no. 4, pp. 288–300, 2010.

[3] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Commun.*, vol. 52, no. 4, pp. 341–353, 2010.

[4] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy.," *Med. Eng. Phys.*, vol. 30, no. 4, pp. 419–425, May 2008.

[5] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proceedings of ICASSP*, Hong Kong, Hong Kong, 2003, vol. 5, pp. 708–711.

[6] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface," in *Proceedings of Interspeech*, Brighton, England, 2009, pp. 640–643.

[7] M. Wand and T. Schultz, "Session-independent EMG-based Speech Recognition.," in *Proceedings of Biosignals*, Rome, Italy, 2011, pp. 295–300.

[8] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-audible murmur (NAM) recognition," *IEICE Trans. Inf. Syst.*, vol. 89, no. 1, pp. 1–8, 2006.

[9] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1189–1197, 2010.

[10] T. Hueber and G. Bailly, "Statistical Conversion of Silent Articulation into Audible Speech using Full-Covariance HMM," Computer Speech and Language, to appear.

[11] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further investigations on EMG-to-speech conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 365–368.

[12] M. Lincoln, A. Packman, and M. Onslow, "Altered auditory feedback and the treatment of stuttering: A review," *J. Fluen. Disord.*, vol. 31, no. 2, pp. 71 – 89, 2006.

[13] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and Yvert, B., "Robust articulatory speech synthesis using deep neural networks for BCI applications.," in *Proceedings of Interspeech*, Singapor, 2014.

[14] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. Part Commun.*, vol. 66, pp. 10–18, 1983.

[15] M. Wand and T. Schultz, "Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation," in *Proc. Interspeech*, 2014, pp. 1189–1193.

[16] T. Hueber, P. Badin, C. Savariaux, C. Vilain, and G. Bailly, "Differences in articulatory strategies between silent, whispered and normal speech? a pilot study using electromagnetic articulography," in *Proceedings of International Seminar on Speech Production (ISSP)*, Montreal, Canada, 2010.

[17] M. Janke, M. Wand, and T. Schultz, "Impact of lack of acoustic feedback in EMG-based silent speech recognition.," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 2686–2689.

[18] "The SPTK toolkit." [Online]. Available: http://sp-tk.sourceforge.net/.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[20] M. Gales and S. Young, "The Application of Hidden Markov Models in Speech Recognition," *Found. Trends® Signal Process.*, vol. 1, no. 3, pp. 195–304, 2007.