Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression

Diandra Fabre, Thomas Hueber, Pierre Badin

GIPSA-Lab / DPC, UMR 5216, CNRS – Grenoble - Alpes University, France

{diandra.fabre, thomas.hueber, pierre.badin}@gipsa-lab.grenoble-inp.fr

Abstract

This paper presents a method for automatically animating the articulatory tongue model of a reference speaker from ultrasound images of the tongue of another speaker. This work is developed in the context of speech therapy based on visual biofeedback, where a speaker is provided with visual information about his/her own articulation. In our approach, the feedback is delivered via an articulatory talking head, which displays the tongue during speech production using augmented reality (*e.g.* transparent skin). The user's tongue movements are captured using ultrasound imaging and parameterized using the PCA-based EigenTongue technique. Extracted features are then converted into control parameters of the articulatory tongue model using Gaussian Mixture Regression. This procedure was evaluated by decoding the converted tongue movements at the phonetic level using an HMM-based decoder trained on the reference speaker's articulatory data. Decoding errors were then manually re-assessed in order to take into account possible phonetic idiosyncrasies (*i.e.* speaker / phoneme specific articulatory strategies). With a system trained on a limited set of 88 VCV sequences, the recognition accuracy at the phonetic level was found to be approximately 70%.

Index Terms: articulatory tongue model, articulatory talking head, ultrasound imaging, GMM, speech therapy

1. Introduction

In order to treat articulation disorders, it could potentially be helpful for both the patient and the therapist to display the position and shape of the tongue in the vocal tract. Besides, several studies have shown that providing a speaker with a visual feedback of his/her own articulation could improve the rehabilitation process (*cf.* [1] for an overview). One of the most widely used techniques is Electropalatography (EPG), which measures timing and location of tongue contact with the hard palate during speech. The use of EPG patterns as a visual feedback tool has been investigated for treating different kinds of articulation disorders, such as those associated with deafness [2] or cleft palate [3]. The use of ultrasound imaging for biofeedback has also been investigated for speech rehabilitation [4] [5]. This non-invasive and clinically safe technique provides a partial view of the tongue during speech and its use in therapy is very promising. However, ultrasound images might prove difficult to interpret for an inexperienced user. Indeed, it is plagued by a typical noise called *speckle*. Besides, it does not show the limits of the oral cavities, *i.e.* neither the palate nor the pharyngeal wall.

Another possible approach to display a target articulatory gesture is to use a so-called *articulatory talking head (ATH)*, *i.e.* a virtual head able to display the internal articulators (tongue, velum) using augmented reality (*e.g.* a transparent skin). Contrary to ultrasound images, an ATH makes the display very intuitive, since it displays all the internal structures of the vocal tract. Several approaches to use an ATH as a feedback tool have been proposed in the context of second language learning. In [6] for instance, pre-recorded animations of an ATH were used to teach Swedish phonemes to French learners. The most appropriate animations were chosen by the experimenter, *i.e.* a phonetically-trained and native speaker of Swedish, who listened to learners' production and selected in response the closest articulatory gesture from a database containing typical errors made by French learners of Swedish.

In our previous work [7, 8], we proposed a system in which the visual feedback was calculated automatically from the user's voice. Different statistical mapping techniques (based on GMM and HMM) were proposed to estimate the most likely articulatory trajectories from the user's acoustics and represent them in the geometrical space of the ATH. This approach gave encouraging results on non-pathological speakers. However, though we have not yet explicitly evaluated it, we assume in the present work that such an approach would be difficult for pathological speakers, since it relies on the speech acoustics only.

Based on these considerations, we developed a system which combines 1) ultrasound imaging in order to capture directly the user's tongue movements rather than estimating it from the acoustics, and 2) an ATH aimed to provide the most intuitive display. As illustrated in Figure 1, Gaussian Mixture Regression is used to convert the visual features extracted from the recorded ultrasound data into a sequence of control

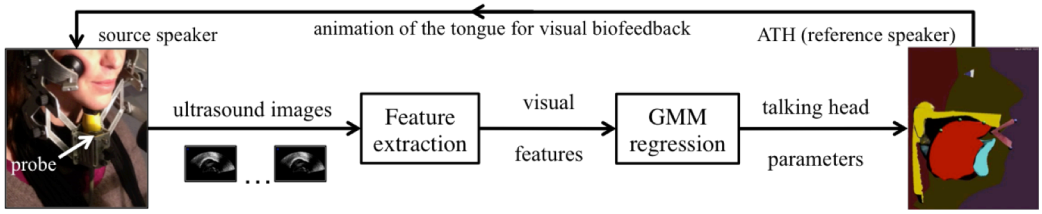


Figure 1. Proposed system of visual feedback based on the automatic animation of an articulatory talking head from ultrasound imaging.

parameters of the ATH. This article presents the technical details of this new technique, by focusing on the animation of the tongue model of the ATH only (the animation of lips, velum and jaw model will be addressed in future work). The proposed technique is evaluated using both automatic and expert-based approaches. In order to discuss how far the proposed system is from a practical clinical application, a special focus is put on the amount of training data that is needed to reach acceptable performance.

Section 2 describes the different steps of the proposed system and details the acquisition and the processing of ultrasound data, the ATH, and finally the GMM-based mapping. Experimental protocol and evaluation procedures are presented in Section 3. Results are finally presented and discussed in the last section.

2. Methodology

2.1. Ultrasound data acquisition and processing

In the proposed approach, tongue movements of the user, referred to as the *source speaker*, are captured using an ultrasound scanner with the probe placed beneath the chin (Figure 1). This probe is positioned in order to obtained tongue images in midsagittal plane and is maintained stationary in relation to the skull using a specific helmet (manufactured by Articulate Instruments ©). The recording of ultrasound images synchronously with the audio signal is achieved using the Ultraspeech system [9] (www.ultraspeech.com). Ultrasound images are recorded at 60 Hz with a resolution of 320x240 pixels, and are then post-filtered using the anisotropic filtering technique [10]. Tongue movements are then encoded using the EigenTongue feature extraction technique [11]. This technique consists in 1) finding the direction of maximum variance in the pixel domain (*i.e.* the EigenTongues) by performing a principal component analysis on a subset of (carefully chosen) training frames – and 2) encoding each image by its N-first projections onto these directions. In this study, we chose to keep 20 visual features (*i.e.* N=20), which represent 85% of the variance observed in the training set.

The spectral content of the audio signal is parameterized by 25 mel-cepstral coefficients (Blackman window, 25ms frame length, 10ms frame shift).

2.2. Articulatory talking head driven by EMA data

The proposed system of visual feedback is based on the articulatory talking head (ATH) developed at GIPSA-lab [12]. This talking head consists of a set of static 3D models of all the speech articulators (*i.e.* lips, tongue, velum, jaw, face) derived from MRI and stereoscopic video images of a speaker referred here to as the *reference speaker*. In [10], we showed that this ATH can be animated from dynamic articulatory data recorded on the same reference speaker using Electromagnetic Articulography (EMA). In the present work, we focused on the animation of the 3D model of the tongue, which corresponds to 6 parameters, *i.e.* the *x* and *y* coordinates of 3 EMA sensors, describing respectively the position of the tip, the dorsum and the back of the tongue. We used a large database of EMA+audio data of the reference speaker, originally described in [13], which consists of two repetitions of 224 VCVs (Vowel Consonant Vowel sequences), two repetitions of 109 pairs of CVC real French words, and 88 sentences (approximately 17 minutes of speech, long pauses being excluded).

2.3. GMM-based mapping between ultrasound images and EMA articulatory data

In the training stage, the source speaker was asked to pronounce a subset of the corpus recorded by the reference speaker. In order to synchronize these movements with those of the reference speaker, the audio signal of source and reference speakers were time-aligned using DTW (Dynamic Time Warping), as schematized in Figure 2.

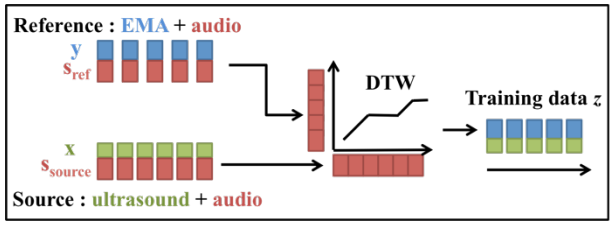


Figure 2. Time-alignment of ultrasound/EMA articulatory data of the tongue recorded on the source/reference speaker using DTW

Finally, the joint probability density function (*pdf*) of source and reference articulatory features (derived respectively from ultrasound and EMA) was modeled by a Gaussian Mixture Model (GMM), such as:

$$p(\mathbf{z} \mid \Theta) = p(\mathbf{x}, \mathbf{y} \mid \Theta) = \sum_{m=1}^{M} \alpha_m N(\mathbf{z}, \mu_m^{\mathbf{z}}, \Sigma_m^{\mathbf{z}})$$
$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \mu_m^{\mathbf{z}} = \begin{bmatrix} \mu_m^{\mathbf{x}} \\ \mu_m^{\mathbf{y}} \end{bmatrix}, \Sigma_m^{\mathbf{z}} = \begin{bmatrix} \Sigma_m^{\mathbf{xx}} \Sigma_m^{\mathbf{yx}} \\ \Sigma_m^{\mathbf{xy}} \Sigma_m^{\mathbf{yy}} \end{bmatrix} \quad (1)$$

where $\mathbf{x}$ is a vector of eigentongue coefficients, $\mathbf{y}$ is a vector of EMA coordinates, Θ is the parameter set of the model, $N(., \mu_m, \Sigma_m)$ is a normal distribution with mean μ and covariance matrix Σ, M is the number of mixture components, and α_m is the weight associated with the m^{th} mixture component (prior probability). Given a training dataset of source and target feature vectors, the *Maximum Likelihood* estimation (ML-estimation) of the GMM parameters Θ_{ML} is determined using the expectation-maximization algorithm (EM) (K-means algorithm is used to obtain an initial clustering of the training set).

In the mapping stage, the target EMA coordinates of the tongue $\hat{\mathbf{y}}_t$ are derived from the ultrasound visual features $\mathbf{x}_t$ such as:

$$\hat{\mathbf{y}}_t = E[\mathbf{y}_t \mid \mathbf{x}_t] = \sum_{m=1}^{M} P(c_m \mid \mathbf{x}_t, \Theta) \cdot \left[\mu_m^{\mathbf{y}} + \Sigma_m^{\mathbf{yx}} \Sigma_m^{\mathbf{xx}^{-1}} (\mathbf{x}_t - \mu_m^{\mathbf{x}}) \right]$$
$$\text{where } P(c_m \mid \mathbf{x}_t, \Theta) = \frac{\alpha_m N(\mathbf{x}_t, \mu_m^{\mathbf{x}}, \Sigma_m^{\mathbf{xx}})}{\sum_{l=1}^{M} \alpha_l N(\mathbf{x}_t, \mu_l^{\mathbf{x}}, \Sigma_l^{\mathbf{xx}})} \quad (2)$$

Finally, the estimated EMA coordinates are used to animate the ATH. Note that we used the MSE estimator and not the MLE estimator (trajectory GMM) [14], which did not bring any improvement.

2.4. Experimental protocol

The proposed technique was evaluated on a French female *source speaker*, with no articulation disorders. She was asked to pronounce a set 110 symmetrical VCV sequences, where V was selected from the French vowels [a ɛ e i y u o ɔ œ ø] and C from the French consonants [t d n ʁ ʃ k g s z l ʒ]. Her tongue movements were recorded using ultrasound imaging,

synchronously with the audio speech signal, and parameterized as described in section 2.1. Note that this corpus is a subset of the EMA+audio corpus recorded by the *reference* speaker (a French male speaker).

Several partitions of the database were considered in order to study how the performance was affected by the size of the training set. This aspect is potentially critical for a clinical use since the amount of required training data has to remain as limited as possible. For that purpose, the database was first randomly divided into 5 parts of equal length. Three evaluation experiments were then conducted. These experiments (referred to as E1, E2, and E3) differ in the amount of data used for training. For each experiment, a cross validation procedure is used as follows:

- Experiment E1: 4/5 of the database were used for training (*i.e.* 88 VCV) and 1/5 for test. (5 possible combinations of training/test sets).
- Experiment E2: 3/5 of the database were used for training (*i.e.* 66 VCV) and 1/5 was used for test (20 possible combinations of training/test sets).
- Experiment E3: 2/5 of the database were used for training (*i.e.* 44 VCV) and 1/5 was used for test (30 possible combinations of training/test sets)

For each experiment, some sequences of the training set were used to estimate the optimal number of mixture components of the GMM, which was found to be 14 for E1 and 12 for both E2 and E3.

Since the source and reference speakers do not have the same vocal tract, it is not possible to calculate directly distances between estimated and measured articulatory tongue movements. Thus, the performances were evaluated using an *articulatory recognition* paradigm, similar to the one used in [8]. An HMM-based phonetic decoder was used to measure the accuracy of the estimated articulatory trajectories at the phonetic level. The decoder was trained on the large database of articulatory data recorded on the reference speaker (described in section 2.2), using a standard training procedure (context-dependent triphone tied-state HMMs). The recognition accuracy defined as $Acc_{art} = (N-D-S-I)/N$ (where N is the total number of phones in the test set, S, D and I are respectively the number of substitution, deletion, and insertion errors) was considered as a measure of the accuracy of the synthetic articulatory trajectory. In order to alleviate the problem of insertion/deletion errors due to the absence of a language model, and to focus on lingual articulations, the decoder was forced to decode either isolated vowels or VCV sequences with C = [t d n ʁ ʃ k g s z l ʒ]. Quite naturally, substitution errors within the following groups: {t d n}, {s z}, {ʃ ʒ}, {k g}, were not considered as errors, as these sets share nearly the same places of articulation.

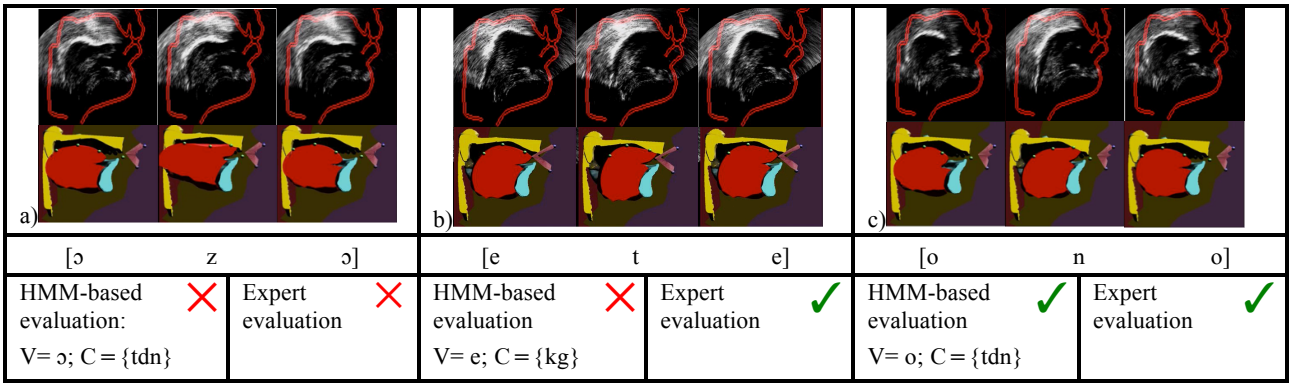
Since the HMM-based phonetic decoder was trained only on the reference speaker's data, it is *a priori* not able to deal with phonetic idiosyncrasy, *i.e.* differences in terms of articulatory strategies between source and reference speaker. Therefore, we supplemented the automatic evaluation with a post re-assessment of the converted tongue movements by expert phoneticians. This procedure can be summarized as follows. Three external experts (speech scientists specialized in speech production) were asked to watch a series of videos showing simultaneously the original ultrasound image sequence of the tongue, and the resulting animation the ATH's tongue model (as shown in Figure 3). In order to help the experts to better interpret the original tongue movement, ultrasound images were supplemented by the contours of the hard palate, upper incisors and pharyngeal wall extracted from an MRI anatomical midsagittal scan of the source speaker's vocal tract. Ultrasound images were then transformed in order to be compatible with the MRI coordinate systems. The rigid transformation parameters (rotation, translation and rescaling) were determined manually by fitting the tongue contours derived from both modalities for the extreme vowels [i] [a] and [u]. The video showing both original and converted tongue movements was finally synchronized with the original audio speech signal.

The experts were asked to evaluate the converted tongue gesture as a whole, with a special focus on the place of articulation of the central consonants, which remains the most challenging issue. In order to limit the length of the test, we asked the expert to re-assess only the sequences for which the central consonant was considered as incorrect by the HMM-based phonetic decoder, for the experiment E1 only. Experts were allowed to watch the videos as many times as they wished, frame-by-frame; they could label the consistency of each consonant between the original and the converted tongue gesture as clearly acceptable, clearly incorrect, or uncertain. Finally, only the tongue movements considered as acceptable by 2 or more experts were re-labeled as "correctly converted".

3. Results and discussion

Table 1 summarizes the recognition accuracies generated by the HMM-based phonetic decoder in experiments E1, E2 and E3. With an accuracy of almost 70% (E1), the proposed method is most of the time able to map correctly the tongue movements of the source speaker into the articulatory space of the ATH (in order to quantify the maximum performance that

Figure 3. Illustration of 3 VCV sequences: one labeled as incorrect by both the phonetic decoder and the experts (a), one labeled as incorrect by the phonetic decoder but correct by the experts (b) and one labeled as correct by both the decoder and the experts (c) (the corresponding video sequences are available in `934_ultrasound_ATH.wmv`)



could have been expected, we also evaluated the performance of the HMM decoder directly on the original articulatory data of the reference speaker and obtained an accuracy of 87 %).

	Total	Vowels	Consonants
E1	69,1 ± 9,9%	71,2 ± 11,9%	63,6 ± 17,7%
E2	68,6 ± 5,0 %	79,1 ± 5,4%	60,3 ± 6,5%
E3	62,4 ± 4,3 %	77,3 ± 4,5 %	48,2 ± 7,6%

Table 1. Articulatory recognition accuracy (with 95% confidence interval) for experiments E1 E2 and E3 (with a training set respectively composed of 88, 66 and 44 VCV sequences).

As shown in Table 1, the performance is almost identical for the two largest training sets, and remains acceptable even with a very limited amount of training data (62.4% in E3 with only 44 VCVs used for training). The sequences labelled as incorrect by the HMM decoder in experiment E1 were then re-assessed by expert phoneticians, following the procedure described in section 2. Figure 3 illustrates the 3 possible situations: the central consonant of a VCV sequence is in 3a) labeled as incorrect by both the phonetic decoder and the pool of experts, in 3c) labeled as correct by both the decoder and the experts. Figure 3b) illustrated the case were the pool of experts re-labeled as correct a sequence considered as incorrect by the decoder: here, the estimated articulatory trajectory for VCV [ete] was decoded as [eke], possibly due to a position of the tongue dorsum too close to the palate during the initial vowel [e]. It was then re-labeled as [ete] by the experts since the place of articulation of the central consonant was compatible with a [t]. After manual re-assessments of the error made on the central consonants, the performance of the proposed approach increases up to 75.1%. In order to analyze these results in greater detail, we present in Table 2 the confusion matrix obtained for lingual consonants in experiment E1 from the HMM decoder, and the matrix obtained when taking into account the manual re-assessment of the errors by the experts.

Table 2. Confusion matrix for lingual consonants generated by the HMM-based phonetic decoder (top); confusion matrix obtained after the manual re-assessment by the experts (bottom) (experiment E1, %c: percentage of correct results, del: deletion)

E1	[tdn]	[sz]	[ʃʒ]	[gk]	[ʁ]	[l]	del	%c
[tdn]	**14**	4	3	2	3	4	0	46,7
[sz]	6	**10**	0	1	0	2	1	50,0
[ʃʒ]	0	0	**16**	1	0	3	0	80,0
[gk]	0	0	0	**19**	1	0	0	95,0
[ʁ]	0	0	0	3	**6**	1	0	60,0
[l]	1	0	4	0	0	**5**	0	50,0

E1	[tdn]	[sz]	[ʃʒ]	[gk]	[ʁ]	[l]	del	%c
[tdn]	**26**	2	1	0	0	1	0	86,7
[sz]	2	**15**	0	1	0	1	1	75,0
[ʃʒ]	0	0	**18**	0	0	2	0	90,0
[gk]	0	0	0	**19**	1	0	0	95,0
[ʁ]	0	0	0	2	**7**	1	0	70,0
[l]	1	0	2	0	0	**7**	0	70,0

Before the manual re-assessment, most of the substitution errors were found for consonants articulated in the alveolar region as [tdn] and [sz]. This might be due to the lack of information on the tongue tip in ultrasound images, as it is often hidden by the acoustic shadow of the mandible. Many of these errors were re-labeled as correct by the experts (46.7% → 86.7% for [tdn] and 50% → 75% for [sz]). However, these results have to be taken cautiously since it is difficult for the experts to evaluate the tiny differences between [tdn] and [sz] in term of place of articulation (for instance between [iti] and [isi]) in either the ultrasound image or the ATH. In that case, they might have relied more on the audio than of the visual modalities (ultrasound and ATH).

Some of the decoding errors re-labeled as "correct" by the experts were also due to phonetic idiosyncrasies of source and reference speakers. This happened in particular for the French fricative consonant [ʁ] which seems to be articulated more backward by the source speaker than by the reference speaker.

4. Conclusions and perspectives

This paper presents a method for automatically animating the tongue model of an articulatory talking head from ultrasound images. Gaussian mixture regression was used to convert visual features extracted from ultrasound images into control parameters of the tongue articulatory model. The accuracy of the estimated movements at the phonetic level was evaluated using an HMM-based articulatory recognizer. Sequences with decoding errors were then manually re-assessed by expert phoneticians in order to take into account possible phonetic idiosyncrasies. With a system trained on a limited set of 88 VCV sequences, the recognition accuracy at the phonetic level was found to be approximately 70% (and even 75% after manual re-assessment). In addition, we found that the performance remained relatively stable when the system was trained with less data.

Future work will focus on the evaluation of the proposed technique on pathological speakers. The most challenging issue will be to deal with mispronunciations during the recording of the training data, which might introduce inconsistency in the model and thus may degrade the general performance. In that context, the use of more advanced mapping techniques, such as Deep Neural Networks (DNN) will be envisioned. In addition, we aim to combine this technique with our previous work on acoustic-to-articulatory inversion by supplementing the ultrasound tongue image with the speech acoustics as input to our system of visual biofeedback.

5. Acknowledgements

The authors would like to thank the Région Rhône-Alpes ARC6 for supporting this work through doctoral funding to Diandra Fabre. They also thank the speaker MG and the different experts for their participation in this work.

6. References

[1] P. Badin, A. Ben Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *L2SW, Workshop on "Second Language Studies: Acquisition, Learning, Education and Technology"*, Tokyo, Japan, 2010, pp. P1-10.

[2] D. W. Massaro and J. Light, "Using visible speech to train perception and production of speech for individuals

with hearing loss," *Journal of Speech, Language, and Hearing Research,* vol. 47, pp. 304-320, 2004.

[3] T. Bressmann, C. Heng, and J. Irish, "Applications of 2D and 3D ultrasound imaging in speech-language pathology," *Journal of Speech Language Pathology and Audiology,* vol. 29, p. 158, 2005.

[4] O. Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher," *Computer Assisted Language Learning,* vol. 25, pp. 37-64, 2012.

[5] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication,* vol. 52, pp. 270-287, 2010.

[6] O. Engwall, "Can audio-visual instructions help learners improve their articulation? — An ultrasound study of short term changes," in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 2631-2634.

[7] T. Hueber, G. Bailly, P. Badin, and F. Elisei, "Speaker adaptation of an acoustic-articulatory inversion model using cascaded Gaussian mixture regressions," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 2753-2757.

[8] T. Hueber, A. Ben Youssef, G. Bailly, P. Badin, and F. Elisei, "Cross-speaker acoustic-to-articulatory inversion using phone-based trajectory HMM," in *Proc. of Interspeech*, Portland, Oregon, USA, 2012, p. Tue.SS3.08.

[9] T. Hueber, G. Chollet, B. Denby, and M. Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in *8th International Seminar on Speech Production, ISSP8*, Strasbourg, France, 2008, pp. 365-368.

[10] Y. Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," *Image Processing, IEEE Transactions on,* vol. 11, pp. 1260-1270, 2002.

[11] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar*, et al.*, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *ICASSP (1)*, 2007, pp. 1245-1248.

[12] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data," in *Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098)*, F. J. Perales and R. B. Fisher, Eds., ed Berlin, Heidelberg, Germany: Springer Verlag, 2008, pp. 132–143.

[13] A. Ben Youssef, P. Badin, G. Bailly, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," in *Proc. of Interspeech*, Brighton, UK, 2009, pp. 2255-2258.

[14] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication,* vol. 50, pp. 215-227, 2008/3 2008.