Can tongue be recovered from face? The answer of data-driven statistical models

Atef Ben Youssef, Pierre Badin, Gérard Bailly

GIPSA-lab (Département Parole & Cognition / ICP), UMR 5216 CNRS – Grenoble University,
961 rue de la Houille Blanche, D.U. - BP 46, F-38402 Saint Martin d'Hères cedex, France
{Atef.BenYoussef, Pierre.Badin, Gerard.Bailly}f@gipsa-lab.grenoble-inp.fr

Abstract

This study revisits the face-to-tongue articulatory inversion problem in speech. We compare the Multi Linear Regression method (MLR) with two more sophisticated methods based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), using the same French corpus of articulatory data acquired by ElectroMagnetoGraphy. GMMs give overall results better than HMMs, but MLR does poorly. GMMs and HMMs maintain the original phonetic class distribution, though with some centralisation effects, effects still much stronger with MLR. A detailed analysis shows that, if the jaw / lips / tongue tip synergy helps recovering front high vowels and coronal consonants, the velars are not recovered at all. It is therefore not possible to recover reliably tongue from face.
Index Terms: EMA, MLR, HMM, articulatory trajectory formation, GMM, HTK, HTS.

1. Introduction

Since more than a decade, the question whether tongue shape can be predicted from lips and face shape is still debated ([1], [2], [3], [4], [5]). So far, these studies were all based on linear modelling. The present study revisits this problem with more sophisticated learning techniques made possible by the availability of large corpora of articulatory data: Hidden Markov Models (HMMs, *cf.* [6]), Gaussian Mixture Models (GMM, *cf.* [7]), and compares the results with those obtained with linear models. The following sections present the state-of-the-art in Face-to-Tongue inversion, the articulatory data used in the study, the three approaches explored, the evaluation of the results, and conclusions.

2. State-of-the-art

All the studies found in the literature used similar articulatory data: one point on the jaw and three points on the tongue recorded by ElectroMagnetoGraphy (EMA), simultaneously with face and lips movements captured by a marker tracking devices (12 or 18 Optotrak points in [1], 17 Qualisys points in [2], 25 Qualisys points in [4] and [5]). By exception, [3] use midsagittal contours traced from X-ray pictures: the tongue is represented by the parameters of a midsagittal articulatory model that fits its shape, while the face and lips are represented by those of another associated 2D model. A tongue model is also used in [4]. Note that tongue and face / lips data in [1] were not acquired simultaneously and had to be time aligned by Dynamic Time Warping (DTW).

The size and nature of the corpus vary a lot: a few sentences repeated 5 times by one AE speaker (total ~400 syllables) and 4 times by one Japanese speaker (total ~400 syllables) in [1]; 69 CV syllables with /a, i, u/ and 23 consonants, and 3 sentences repeated 4 times (total ~520 syllables) by four AE speakers in [2]; 45 frames selected at the centre of VCV syllables produced by one French speaker in [3]; 63 VCV with /a i u/ context uttered once by one Swedish speaker in [4]; 138 symmetric VC1{C2C3}VCV, 41 CVC and 270 short sentences (total ~2500 syllables) uttered by one Swedish speaker in [5].

All studies use Multi Linear Regression (MLR) to predict tongue data from face data. The inversion is assessed by computing the *Pearson product-moment correlation coefficient* (PMCC) between measured and predicted data. In a jack-knife training procedure, the data are split into *n* parts of which *(n -1)* are used to determine the MLR coefficients, and to predict the *n*-th remaining part. The PMCC coefficient is the average over the *n* values of the correlation coefficients between obtained by the jack-knife procedure. The factor *n* is set to 4 or 5 in [1] and [2], to 1 in [3], and to 10 in [4] and [5].

Results are summarised in Table 1. The first line refers to tongue coils receptors: (*Tx*, *Ty*), (*Mx*, *My*) and (*Bx*, *By*) correspond to the horizontal and vertical midsagittal coordinates of the coils attached respectively to the tongue tip, the tongue middle, and the tongue back; moreover, *G* refers to the PMCC computed over the six coordinates. For [3] and [4], *TB*, *TD*, *TT* and *TA* (light blue in Table 1) refer respectively to the tongue *Body*, *Dorsum*, *Tip* and *Advance* control parameters of the articulatory tongue model. From their results, [1] claim that the tongue can be recovered reasonably well from facial motion; however, if we exclude jaw and lips coils from their predicted data, we find only medium correlations (0.65 – 0.79). Medium to high correlations are found in [2], whereas lower correlations are obtained by [4], and also by [3] and [4] when using an articulatory model to track speech movements. On a larger corpus, [5] gets a still lower global correlation.

Interestingly, tongue tip (either *Ty* or *TT*) appears to be the tongue region best recovered in all studies: [3] suggests that this may be ascribed to the fact that the jaw is an articulator with a strong influence on both labial and lingual shapes.

Phonetic context has a clear influence on the results: [2] and [3] note that results are better for C/a/ syllables than for C/i/ and C/u/ syllables, while [4] describes a more complex pattern. Bailly & Badin [3] remark that articulations

Table 1. *Correlation coefficients for each EMA coordinate (**bold** for maximum and italics for minimum values) for the various studies.*

	Tx	Mx	Bx	Ty	My	By	G
[1]	0.66	0.66	**0.71**	0.68	*0.57*	0.60	0.65
	0.81	**0.83**	**0.83**	0.76	0.80	*0.72*	0.79
[2]	0.72	*0.69*	0.71				**0.74**
	0.80	**0.85**	**0.85**				**0.83**
[5]							**0.52**
[4]	**0.83**	0.72	0.68	**0.83**	*0.35*	0.80	**0.66**
	TA	TB	TD	TT			
	0.26	0.54	0.40	**0.75**			0.49
[3]	*0.37*	0.71	0.64	**0.74**			0.62
MLR	0,58	0,61	0,58	0,78	0,55	0,39	**0.59**
HMM	0.71	0.70	0.72	0.79	0.68	0.55	**0.70**
GMM	0,83	0,82	0,80	0,87	0,81	0,63	**0.80**

associating a jaw/tongue/lips synergy along the axis closed/front (*e.g.* [i]) vs. open/back (*e.g.* [a]) are more accurately recovered than those requiring constrictions deviating from this synergy. In complement, [4] note that face information is insufficient to accurately predict a non alveolar vocal tract constriction, which is in line with [3].

The fact that the lowest mean correlation is obtained in the study with the largest corpus ([5]), in complement to the fact that correlations are higher for CVs in context than for sentences ([2]) suggests that linear methods may be efficient for restricted ranges of articulations, but less able to cope with the full range of speech movements.

3. Articulatory data

The three methods for Face-to-Tongue inversion that we have explored are based on the articulatory data already presented and used by [8] for acoustic-to-articulatory inversion, and by [9] for tongue reading experiments. This section gives a brief overview.

The corpus uttered by a French male speaker consists of a set of VCV nonsense sequences, CVC real words, and sentences. At first, the audio signal was used to label the allophones in each utterance, using the corresponding phonetic transcription string, by a forced alignment procedure based on HMMs. The allophones centres were automatically chosen as the average between beginning and end of the phonemes. The 36 phonemes were: [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃ p t k f s ʃ b d g v z ʒ m n ʁ l w ɥ j ə _ __], where _ and __ are internal short and utterance initial and final long pauses respectively. The corpus, from which long pauses were excluded, contained finally about 100,000 frames (~17 mn), corresponding to 5132 phones and about 2500 syllables.

The articulatory data have been recorded by means of an ElectroMagnetic Articulograph (EMA) that allows tracking flesh points of the articulators thanks to small electromagnetic receiver coils. Studies have shown that the number of degrees of freedom of articulators (jaw, lips, tongue, …) for speech is limited, and that a small number of carefully selected measurement locations can allow retrieving them with a good accuracy ([9]). In the present study, six coils were used: a jaw coil attached to the lower incisors (midsagittal coordinates: *Jx*, *Jy*); an upper lip coil (*ULx*, *ULy*) and a lower lip coil (*LLx*, *LLy*) attached to the boundaries between the vermilion and the skin; three coils attached to the tongue tip (*Tx*, *Ty*), middle (*Mx*, *My*), and back (*Bx*, *By*).Note that the EMA coordinates were low passed at 20 Hz, and down sampled at 100 Hz.

4. Three inversion methods

In this study, we considered that the face data consisted of the lip coils coordinates, complemented by the jaw coil coordinates, as Revéret *et al.* [10] have shown that jaw height can be predicted from face points. The tongue data consisted of the three tongue coils coordinates.

4.1. Multi Linear Regression modelling

Following the previous studies described above, we have first modelled the relations between face and tongue coordinates by a. Multi Linear Regression (MLR) model. MLR allows finding the matrix $\boldsymbol{A}$ that ensures the optimal fit, *i.e.* the minimal Root Mean Square Error (RMSE) between measured and modelled parameters, as:

$$\hat{Y}_{tT} = A \times Y_{tF} \quad (1),$$

where $\boldsymbol{Y_{tF}}(1{:}N_t, 1{:}n_F)$ is the matrix of the $n_F = 6$ measured face coils coordinates [*Jx*, *Jy*, *ULx*, *ULy*, *LLx*, *LLy*] for the N_t time instants of the *testing* set, and $\boldsymbol{\hat{Y}_{tT}}(1{:}N_t, 1{:}n_T)$ is the matrix of the $n_T = 6$ tongue coils coordinates [*Tx*, *Ty*, *Mx*, *My*, *Bx*, *By*] estimated for the *testing* set. The linear model matrix $\boldsymbol{A}(1{:}n_F, 1{:}n_T)$ is classically computed over the *training* set as:

$$\mathrm{A} = (\mathrm{Y_F Y_F}^T)^{-1} \mathrm{Y_F Y_T}^T \quad (2),$$

where $\boldsymbol{Y_T}(1{:}N, 1{:}n_T)$ and $\boldsymbol{Y_F}(1{:}N, 1{:}n_T)$ are the measured tongue and face coordinates for the N time instants of the *training* set. The errors between $\boldsymbol{\hat{Y}_{tT}}$ and $\boldsymbol{Y_{tT}}$ are presented in section 5.

4.2. HMMs

The present HMM modelling of speech production is similar to that performed by [8]. For the HMMs training, we considered the face and tongue features vectors (coils coordinates and first time derivatives) as two streams in the HTK multi-stream training procedure. Subsequently, the HMMs obtained were split into *face HMMs* and *tongue HMMs*. Following [8], various contextual schemes were tested: phonemes without context ('*no*'), with left ('*L*') or right ('*R*') context, and with both left and right contexts ('*L-R*').

Left-to-right, 3-state phoneme HMMs with one Gaussian per state and a diagonal covariance matrix were used. For training and test the HTK3.4 toolkit is used [11]. The training was performed using the Expectation Maximization (EM) algorithm based on the Maximum Likelihood (ML) criterion.

A bigram language model considering sequences of phones in context was trained over the whole corpus.

The face-to-tongue inversion is performed in two steps. The first step performs phoneme recognition, based on the *face HMMs*. The result is a sequence of recognised allophones together with the durations of each state of each allophone.

The second step of the inversion aims at reconstructing the tongue articulatory trajectories from the chain of phoneme labels and boundaries delivered by the recognition procedure. As described in [12], the synthesis is performed in two phases: a linear sequence of HMM states is built by concatenating the corresponding phone HMMs, and then a sequence of observation parameters is generated using a specific ML-based parameter generation algorithm, using the software developed by the HTS group ([13]). Note that the state durations were not estimated by z-scoring as in [8], but by the recognition stage.

Due to the limited size of the training sets, some phonemes in context were missing. In order to overcome this problem, an inheritance mechanism is used: each missing *L-R* model is replaced by the corresponding *R* model if it exists and by the context-independent model if this latter model does not exist either.

4.3. GMMs

The GMM was trained using the expectation–maximization (EM) algorithm using the joint face-tongue vectors as training set. The GMM-based mapping is then applied using the minimum mean-square error (MMSE) criterion, which has been often used for voice conversion [14] or in acoustic-to-articulatory inversion ([7]). Moreover, to improve the mapping performance, the maximum likelihood estimation (MLE) was applied to the GMM-based mapping method. The determination of a target parameter trajectory with appropriate static and dynamic properties is obtained by combining local estimates of the mean and variance for each frame $p(t)$ and its derivative $\Delta p(t)$ with the explicit relationship between static and dynamic features (*e.g.* $\Delta p(t) = p(t) - p(t\text{-}1)$) in the MLE-

based mapping. At each time instant indexed by *j*, the feature vector is the concatenation of a variable number '2*n*+1' of vectors of EMA coils coordinates [PCA($\boldsymbol{Y_F}(J, 1{:}n_F)$); $\boldsymbol{Y_T}(j, 1{:}n_T)$; $\boldsymbol{\Delta Y_T}(j, 1{:}n_T)$], where $\boldsymbol{\Delta}$ denotes first time derivation, and $J = j+[-n{:}+n]$ denotes the time instant indices of the set of input frames used for contextual information. The number of input frames was varied from phoneme size (n=4, ~90 ms) to diphone size (*n*=8, ~170 ms), but the dimension '$(2n+1)\times n_F$' of the resulting vector was reduced to a fixed value of 24 by Principal Component Analysis (PCA). The number of mixture components was varied from 16 to 128.

5. Evaluation

Two criteria have been used to assess the inversion results: the RMSE between the measured and recovered coordinates, and PMCC, a less conservative criterion that measures only the level of amplitude similarity and of synchrony of the trajectories. Unless otherwise specified, a jack-knife training procedure was used, splitting the data into five parts of which four were used for training and the remaining one used for testing. The RMSE and PMCC were calculated over the five testing partitions – therefore the whole corpus –, excluding the long pauses at the beginning and the end of each utterance. In addition phoneme recognition rates were also used to assess the phoneme recognition step.

5.1. Evaluation of the HMM-based inversion

Table 2 that displays the RMSE and the PMCC for the HMM-based inversion shows that the best results are obtained for phones with both right and left contexts. We also found that the use of state durations produced by the recognition stage permitted an improvement of about 4 % for both RMSE and PMCC, compared to the previously used z-scoring method. Besides, in order to assess the contribution of the trajectory formation to errors for the complete inversion procedure, we also synthesised these trajectories directly from the original labels, simulating a perfect face recognition step: from Table 2, we can estimate that the contribution of the trajectory formation step to the overall RMSE amounts to about 60 % on average; note that it was nearly 90 % for acoustics to vocal tract articulation inversion experiments with the same methods on the same corpus in a way similar to [8]. This shows that recognition from face is much less efficient than recognition from acoustics. This is confirmed by the results given in Table 3 which shows that – as expected – the performance of face recognition is much lower than that of acoustic recognition, by 30 % on average.

Table 2. *RMSE (mm) and PMCC for the HMM inversion with different types of contexts.*

	Phones from face				Original phones			
Ctxt	**no**	**L**	**R**	**L-R**	**no**	**L**	**R**	**L-R**
RMSE	4,22	3,68	3,67	**3,64**	2,74	2,23	2,17	**1,71**
PMCC	0,55	0,68	0,68	**0,70**	0,85	0,89	0,9	**0,94**

Table 3. *Recognition rates (Percent **Cor**rect, **Acc**uracy) for phoneme recognition from Face and phoneme recognition from Acoustics.*

Ctxt	**no**		**L**		**R**		**L-R**	
	Cor	Acc	Cor	Acc	Cor	Acc	Cor	Acc
Face	58.91	47.86	71.28	46.93	71.03	44.41	**69.46**	**53.71**
Acoust.	88.90	68.99	92.61	78.14	**93.66**	**80.90**	87.12	80.83

5.2. Evaluation of the GMM-based inversion

Table 4 shows the RMSE and PMCC for experiments using different numbers of mixtures and context window sizes. The RMSE decreases when the number of mixtures increases. For 128 mixtures, the optimal context window size is 110 ms. The most plausible interpretation is that a phoneme-sized window optimally contains necessary local phonetic cues for inversion. Using the extra MLE optimisation stage was found to improve the results by 5 %.

Table 4. *RMSE (mm) and PMCC for the GMM inversion (MLE) with different numbers of mixtures (# mix) and context window sizes (ctw).*

#mix	**16**		**32**		**64**		**128**	
ctw	**RMSE**	**PMCC**	**RMSE**	**PMCC**	**RMSE**	**PMCC**	**RMSE**	**PMCC**
90	3.49	0.70	3.20	0.75	3.06	0.78	2.93	0.80
110	3.44	0.71	3.19	0.75	3.02	0.78	**2.90**	**0.80**
130	3.47	0.70	3.19	0.75	3.04	0.78	2.94	0.80
150	3.46	0.70	3.18	0.75	3.03	0.78	2.95	0.79
170	3.49	0.69	3.18	0.75	2.98	0.79	3.27	0.75

5.3. Evaluation of the MLR-based inversion

The inversion based on the MLR model led to an RMSE of 3.88 mm and a PMCC of 0.59, using the jack-knife evaluation procedure. In order to compare our results to those of the other studies, we made complementary experiments on reduced speech material: using one repetition of the symmetrical VCV, where C is one of the 16 French consonants and V = /i a u/ for training and the other repetition for testing, the RMSE was 3.29 mm and the PMCC 0.84, which is comparable to the other studies. Interestingly, when adding the /y/ vowel – which possesses the same lip rounding feature than /u/ in French – to the /a i u/ set, the RMSE rises to 3.67 mm and PMCC decreases to 0.77, which confirms the difficulty to predict the tongue shape from the face shape for a number of articulations.

6. Discussion

This study has shown that the inversion methods based on HMM, GMM and MLR models give RMSE levels of 3.64, 2.90 and 3.88 mm respectively, and correlations of 0.70, 0.80 and 0.59. In order to set a reference for these results, we have also computed (using the jack-knife method) the RMSE restricted to the three tongue coils for the acoustic-to-articulatory inversion using a similar approach (*cf.* [8] for the HMMs): the results were much better with the HMMs (RMSE: 2.22 mm, PMCC: 0.89), which was expected, but a bit worse with the GMMs (2.55 mm / 0.86), which is surprising and unexplained. Table 5 shows that vowels /i a/ are rather well reconstructed with all three methods, while /y u/ are not. Note however the surprisingly good result for /u/ with HMMs, likely due to context effects. Note also that, if the coronal consonant /t/ is well recovered, the velar one /k/ is not. This illustrates the general tendency that coronals are relatively well estimated, while velars are much less, in line with [4].

Visual comparisons of the spaces covered by the coils recovered with those covered by the measured ones have

Table 5. *RMSE for individual phonemes (mm).*

	i	a	u	y	p	t	k
GMM	2.21	2.44	2.98	3.95	2.77	1.76	4.81
HMM	2.85	2.85	4.03	4.46	3.59	2.54	5.25
MLR	3.42	2.88	3.91	5.77	3.72	2.81	5.50

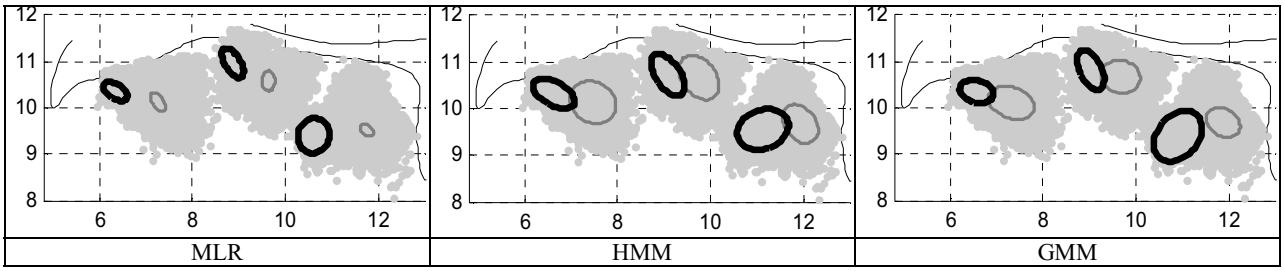

Figure 1: *Dispersion ellipses of coils for phones with errors larger than 10 mm for at least one coil coordinate: original data (thick lines), estimated data (thin lines), superposed on original data points for all phones (light grey dots). Note the general back of the estimates.*

revealed a very strong tendency for MLR to centralise the articulations. HMMs maintain spaces very close to the originals ones, while GMMs induce a small retraction of these spaces; this is a bit surprising, as the RMSE and PMCC estimations rank the GMMs before the HMMs. Figure 1 illustrates this general centralisation tendency for the phones having a recovery error larger than 10 mm for at least one of the six tongue coils coordinates (138, 221 and 71 phones for MLR, HMM and GMM respectively). The light grey background corresponds to the space covered by the original 5132 phones; the ellipses that represent the recovered phones with high errors (thin lines) are much closer to the centres of the corresponding originals spaces (light grey) than the ellipses that represent the corresponding original phones (thick lines). This illustrates the difficulty to predict important characteristics of tongue shape from face shape.

In another attempt to analyse and interpret the results, we have constructed dendrograms, considering only the central frame of each phone, separating vowels and consonants. This was done based on the Mahalanobis distance between each phone class, using Matlab™ functions based on one-way multivariate analysis of variance. Three classes were imposed for the vowels and nine for the consonants. The observation of the dendrograms has shown that: (1) for MLR, the classes for the *predicted tongues* are identical to those for the *measured faces*, which points to an erroneous recovery; (2) for HMM and GMM, the classes for the predicted tongues are identical to those for the measured tongues (with one exception for the vowels with the GMM), but with much lower distances (a dendrogram distance of 3 would have collapsed the consonants in 2 or 3 classes for the predicted tongues, leaving intact the 9 classes for the measured one), which also points to a low reliability of the inversion.

7. Conclusions

We have revisited the Face-to-Tongue inversion problem in speech. Using a much larger corpus than previously in the literature (except for [5]), we have assessed methods of different complexity and found that GMMs gave overall results better than HMMs, and that MLR did poorly. GMMs and HMMs can maintain the original phonetic class distribution, though with some centralisation effects that are still much stronger with MLR. Previous studies gave fairly good overall results, presumably because MRL cope well with limited material: we have shown that for larger corpuses, MLR gives poor results. As suggested by [2], more sophisticated context-sensitive techniques have improved the results fairly much. However, a detailed analysis has shown that, if the jaw / lips / tongue tip synergy helps recovering front high vowels and coronal consonants, the velars are not recovered at all. In conclusion, it is not possible to recover reliably tongue from face.

8. Acknowledgements

Ch. Savariaux for the EMA recordings, and Tomoki Toda (NAIST, Japan) for his GMM toolbox. This work was partially supported by the French ANR-08-EMER-001-02 *ARTIS* and the French-Japanese PHC SAKURA *CASSIS* projects.

9. References

[1] Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E., "Quantitative association of vocal-tract and facial behavior," *Speech Comm.*, vol. 26, pp. 23-43, 1998.

[2] Jiang, J., Alwan, A.A., Keating, P.A., Auer, E.T., Jr, and Bernstein, J., "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP-JASP*, vol. 2002, pp. 1174-1188, 2002.

[3] Bailly, G. and Badin, P., "Seeing tongue movements from outside," Interspeech, Denver, Colorado, USA, 2002.

[4] Engwall, O. and Beskow, J., "Resynthesis of 3D tongue movements from facial data," Eurospeech, Geneva, Switzerland, 2003.

[5] Beskow, J., Engwall, O., and Granström, B., "Resynthesis of facial and intraoral articulation from simultaneous measurements," 15th ICPhS, Barcelona, Spain, 2003.

[6] Zhang, L. and Renals, S., "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245-248, 2008.

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

[8] Ben Youssef, A., Badin, P., Bailly, G., and Heracleous, P., "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme HMMs," Interspeech 2009, Brighton, UK, 2009.

[9] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G., "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Comm.*, vol. 52, pp. 493-503, 2010.

[10] Revéret, L., Bailly, G., and Badin, P., "MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," 6th ICSLP, Beijing, China, 2000.

[11] Young, S., Evermann, G., *et al.*, "The HTK Book. Revised for HTK Version 3.4 December 2006," 2006.

[12] Govokhina, O., Bailly, G., Breton, G., and Bagshaw, P., "TDA: A new trainable trajectory formation system for facial animation," Interspeech, Pittsburgh, PE, 2006.

[13] Zen, H., Tokuda, K., and Kitamura, T., "An introduction of trajectory model into HMM-based speech synthesis," ISCA SSW5, Pittsburgh, PA, USA, 2004.

[14] Stylianou, Y., Cappé, O., and Moulines, E., "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 131-142, 1998.