INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An HMM/DNN Comparison for Synchronized Text-to-speech and Tongue Motion Synthesis

Sébastien Le Maguer[1,2], *Ingmar Steiner*[1−3], *Alexander Hewer*[1−3]

[1]Computational Linguistics & Phonetics, Saarland University, Germany
[2]Multimodal Computing and Interaction, Saarland University, Germany
[3]German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
{slemaguer, steiner, hewer}@coli.uni-saarland.de

Abstract

We present an end-to-end text-to-speech (TTS) synthesis system that generates audio and synchronized tongue motion directly from text. This is achieved by adapting a statistical shape space model of the tongue surface to an articulatory speech corpus and training a speech synthesis system directly on the tongue model parameter weights. We focus our analysis on the application of two standard methodologies, based on Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs), respectively, to train both acoustic models and the tongue model parameter weights. We evaluate both methodologies at every step by comparing the predicted articulatory movements against the reference data. The results show that even with less than 2h of data, DNNs already outperform HMMs.

Index Terms: Text-to-speech, multimodal synthesis, tongue modeling, articulatory animation

1. Introduction

Multimodal speech synthesis which integrates intra-oral articulatory movements has been investigated for several years. These movements are particularly useful for computer-assisted pronunciation training (CAPT) applications and articulatory visualization. Previous studies [e.g., 1, 2], combined intra-oral motion capture data obtained from electromagnetic articulography (EMA) [3] with concatenative speech synthesis to animate a geometric tongue model simultaneously with synthesized audio. Among more recent implementations, the statistical parametric speech synthesis paradigm introduces greater flexibility in the modeling and therefore in the combination of multiple modalities. Consequently, several studies [4, 5, 6] have successfully used hidden Markov model (HMM) based multimodal speech synthesis with EMA data.

In this paper, we present an approach to multimodal text-to-speech (TTS) synthesis that generates the fully animated, three-dimensional (3D) surface of the tongue, synchronized with synthetic audio. This is achieved using data from a single-speaker, articulatory corpus that includes EMA motion capture of three tongue fleshpoints [7]. In contrast to other work, our approach employs a tongue model which can easily be adapted to different speakers.

Such geometrical tongue models have been successfully used in previous work to generate animations from provided articulatory data: Katz et al. [8] presented a real-time visual feedback system that deforms a generic tongue model by using EMA data. However, due to the generic model, their approach did not take anatomical differences into account. A statistical model was used in the approach by Badin et al. [9]. They used the data of one speaker to derive the tongue model and used the EMA data of the same speaker to animate it. Engwall [10] followed a similar approach.

Our own previous work utilized a multilinear statistical model to visualize EMA data, which allowed it to be adapted to different speakers [11]. This work was extended into a full TTS system, where the audio and articulatory motion are synthesized using an HMM based TTS framework [12], while the surface restoration is performed by means of a multilinear statistical tongue model [13] trained on a multi-speaker, volumetric magnetic resonance imaging (MRI) dataset [14].

Advancing statistical parametric TTS synthesis, Deep Neural Networks (DNNs) have been applied with great success [15]. However, the amount of data needed to train a DNN model is quite significant.

Therefore, the present study focuses on the comparison of HMM and DNN modeling to achieve a multimodal speech synthesis system. We are also focusing on the fact that the same amount of data, which is less than 2 h of speech, is used to train the models. Therefore, we also want to compare the behavior of DNN models with HMM based on this relatively limited amount of data.

This paper is organized as followed. First, in Section 2, the architecture of the proposed framework is described. Then, in Section 3, we focus the experiment description and the results analysis. Finally we will conclude this paper.

2. Method

2.1. Multilinear shape space model

In our framework, we employ a multilinear model to generate tongue shapes. Specifically, we use this model to construct a function $f : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathscr{M}$ that maps the parameters $s \in \mathbb{R}^m$ and $p \in \mathbb{R}^n$ to a polygon mesh $M = (V,F) \in \mathscr{M}$. Here, the set $V := \{v_i\}$ is called the vertex set of the mesh with $v_i \in \mathbb{R}^3$ and F is the face set that uses these vertices to describe the represented surface. We note that meshes M have the same face set and only differ in the positional data of their vertices. In our case, the speaker parameter s represents the anatomical features of the generated tongue while the pose parameter p determines its articulation-related shape.

To summarize our previous approach [13], we derive this model in the following manner. We first use image segmentation and template matching methods to extract tongue meshes M from a database that contains MRI recordings of m speakers showing the vocal tract configuration for n different phonemes. Thus, in the end, we have for each speaker one tongue mesh for each considered phoneme available, which we use to derive the degrees of freedom (DoF) of the anatomy and speech related variations as follows: we center the meshes and turn the

http://dx.doi.org/10.21437/Interspeech.2017-936

them into feature vectors by serializing the positional data of their vertices. These feature vectors are then arranged in a third order tensor A, such that its first mode corresponds to the speakers, the second to the considered phonemes, and the third to the positional data. Finally, we use higher order singular value decomposition (HOSVD) [16] to get access to the following tensor decomposition:

$$A = C \times_1 U_1 \times_2 U_2 \tag{1}$$

Here, C is a third order tensor that represents our multilinear model. The operation $C \times_n U$ is the n-th mode multiplication of the tensor C with the matrix U. The rows of $U_1 \in \mathbb{R}^{m \times m}$ correspond to the parameters of the original speakers, the ones of $U_2 \in \mathbb{R}^{n \times n}$ to the parameters of the considered phonemes. We note that in contrast to a principal component analysis (PCA) model, the multilinear model captures anatomical and articulation related shape variations separately. We can use C to generate new positional data for provided weights s and p:

$$v(s,p) = \mu + C \times_1 s \times_2 p \tag{2}$$

where μ is a feature vector consisting of the positional data corresponding to the mean mesh of the tongue shape collection. Finally, we reconstruct the vertex set by using the generated positional data and combine it with the face set to obtain our mesh.

In order to fit the model to a speaker of an EMA corpus and to obtain the associated weights, we apply the following approach. First, we manually align the EMA data to the model space by using a provided reference coil. Then, we find correspondences between the considered tongue coils and vertices on the model mesh in a semi-supervised way: we start with random parameters and generate the associated mesh. Afterwards, we find for each coil the nearest vertex of the mesh. These correspondences are then iteratively refined by fitting the model and updating the nearest vertices. The mentioned three steps are repeated multiple times and those correspondences are kept which result in the smallest average distance between coils and vertices. In the end, the final correspondences are manually inspected and the experiment is repeated if they are deemed to be incorrect. After the correspondences have been obtained, we fit the model to each EMA data frame of the corpus by minimizing the following energy:

$$E(s,p) = E_{\text{Data}}(s,p) + E_{\text{Smooth}}(a,p) \tag{3}$$

The data term $E_{\text{Data}}(\cdot)$ measures the distances between the coil positions and the corresponding mesh vertices of the generated mesh $f(s,p)$. The smoothness term $E_{\text{Smooth}}(\cdot)$ penalizes differences between the current parameters and the ones of the previous time step. We note that the fitting can also be performed while only optimizing for one weight and leaving the other one fixed.

2.2. Multimodal statistical parametric speech synthesis

In this study, we plan to analyze the influence of the Feed Forward Deep Neural Network (FF-DNN) modeling compared to Gaussian Mixture Model (GMM)/decision tree modeling used in the default HMM based synthesis (HTS) system, focusing on the accuracy of the obtained synthesis. To simplify the notations, we use the term DNN to refer to the FF-DNN modeling and the term HMM for the GMM/decision tree modeling. Furthermore, to achieve the comparison, we have adapted the standard architecture of the HTS framework [17] to obtain the one shown in Figure 1.

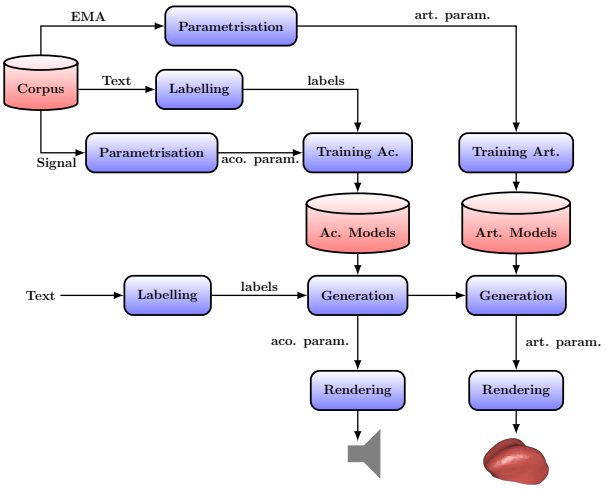

Figure 1: *Adapted architecture for multimodal synthesis.*

The architecture consists of the following main parts: the parameterization of the signal, the training of the models (acoustic/articulatory), the parameter generation (acoustic/articulatory), and the rendering (acoustic/articulatory).

Considering the parameterization of the acoustic signal, we used the standard configuration described by Zen and Toda [17] based on STRAIGHT [18] and the mel log spectrum approximation (MLSA) filter [19]. First, STRAIGHT is used to extract the spectrum, the fundamental frequency (F_0), and the aperiodicity. The F_0 values are transformed into the logarithmic domain, to be more consistent with human hearing. The F_0 trajectory is interpolated and the voiced/unvoiced property is extracted in order to respect the standard DNN training proposed by Zen et al. [15]. Finally, the MLSA filter is used to parameterize these coefficients and to obtain the mel-generalized cepstral coefficients (MGC) and the aperiodicity per band (BAP), respectively.

In addition to the acoustic signal, we introduced the articulatory trajectory parameterization part. In the present study, we work towards replacing the EMA data by the tongue model parameters. Therefore, our goal is to train our models on the trajectories of the tongue model parameters using both modeling methodologies.

To increase flexibility, we separated the training of the acoustic and articulatory models. However, the training stages are either based on the the standard HTS training stage proposed by Zen and Toda [17], or on the standard DNN training proposed by Zen et al. [15].

The standard parameter generation algorithms (described by Tokuda et al. [20]) are then used to obtain the trajectories. More importantly, we want to retain the synchronization between the acoustic and the articulatory trajectories. To do so, the durations produced by the acoustic generation stage are imposed onto the articulatory generation stage at the phone level.

3. Experiments

3.1. Experimental setup

The data used for the experiments in this study is taken from the *mngu0* corpus. Specifically, we used the "day 1" EMA subset [7], which contains acoustic recordings, time-aligned pho-

Table 1: *EMA coil labels and locations of the* mngu0 *corpus.*

label	location
T1	tongue tip
T2	tongue body
T3	tongue dorsum
ulip	upper lip
llip	lower lip
ref	upper incisor
jaw	lower incisor

netic transcriptions, and EMA motion capture data (sampled at 200 Hz using a Carstens AG500 articulograph).

We selected the "basic" (as opposed to the "normalized") release variant of the EMA data. It also preserves the silent (i.e., non-speech) intervals, as well as the 3D nature and true spatial coordinates of the sensor data (after head motion compensation). The EMA coils are listed in Table 1.

From the provided acoustic data, signal parameters were extracted using STRAIGHT [18] with a frame rate of 200 Hz, matching that of the EMA data. As we follow the standard HTS methodology, we also kept the same parameters. Therefore, our signal parameters are 50 MGC, 25 BAP and one F_0 coefficient.

To derive the multilinear model, we used the whole Ultrax dataset [14] (11 speakers) enriched with the data of Baker [21] (1 speaker). We then extracted the pose parameters from the EMA training data as follows. First, we estimated the vertex-coil correspondences by using the upper incisor coil as a reference. Afterwards, we fitted the model to all frames of the EMA data to obtain the speaker and pose parameters and averaged the resulting speaker parameters to estimate the subject's anatomical features. Finally, we again fitted the model to all frames while fixing the speaker parameter to the obtained estimate.

From the 1354 utterances in the data, 152 (11.20 %) were randomly selected and held back as a test set. The remaining 1202 utterances were used as the training set to build the synthesis models.

Considering the training configurations, the default HTS 2.3 setup [22] was used for the HMM voice and the default HTS 2.3.1 setup[1] for the DNN voice. Therefore, the DNN configuration is implying 3 hidden layers containing 1024 nodes each. For both configurations, we adapted the F_0 limits to the interval 60 Hz to 300 Hz.

3.2. Discussion

To achieve the analysis, we have compared, for both conditions, 5 different setups:

straight-ema which is the joint modeling of the acoustic and the EMA features in the same vector;
straight which predicts only the acoustic parameters;
ema which predicts only the articulatory parameters;
ema-tongue which predicts only the articulatory parameters associated with the tongue;
tm which predicts the tongue model parameters (p from Equation (2)).

To compare the different setups, we used the classical acoustic objective distance evaluation composed by the mel cepstral distortion (MCD) for the spectrum part, root mean square error (RMSE) for duration at the phone level, and for F_0 at the frame level, and finally the voiced-unvoiced (VUV) error rate. Considering the articulatory features, we used the Euclidean distance.

[1] `http://hts.sp.nitech.ac.jp/?Download#f2602aa9`

Table 2: *HMM synthesis evaluation results for* straight-ema *(acoustic combined with EMA synthesis).*

id	mean	std. dev.	conf. int.
rms F_0 (cent)	188.43	63.70	10.21
rms F_0 (Hz)	10.66	4.91	0.79
vuv (%)	12.14	3.84	0.62
mcdist (dB)	2.45	0.23	0.04
rms dur. (ms)	41.93	19.04	3.05
eucdist T3 (mm)	2.14	1.47	8.57×10^{-3}
eucdist T2 (mm)	2.10	1.54	9.00×10^{-3}
eucdist T1 (mm)	2.17	1.62	9.44×10^{-3}
eucdist ref (mm)	0.22	0.12	6.97×10^{-4}
eucdist jaw (mm)	1.26	0.65	3.80×10^{-3}
eucdist ulip (mm)	0.72	0.38	2.21×10^{-3}
eucdist llip (mm)	1.45	0.93	5.45×10^{-3}

Table 3: *HMM synthesis evaluation results for* straight *(acoustic synthesis).*

id	mean	std. dev.	conf. int.
rms F_0 (cent)	188.52	76.92	12.33
rms F_0 (Hz)	10.77	5.47	0.88
vuv (%)	12.03	3.94	0.63
mcdist (dB)	2.45	0.22	0.04
rms dur. (ms)	42.00	18.29	2.93

Tables 2, 3 and 4 present the results for the conditions *straight-ema*, *straight* and *tm*, respectively, in the HMM setup. By comparing the setup *straight-ema* and *straight*, we can conclude that the duration modeling is equivalent. Therefore, we can separate the acoustic from the articulatory modeling. Furthermore, the comparison of the Euclidean distances between the *tm* and *straight-ema* setups shows a small degradation. We assume that a combination of measurement noise and reconstruction error (which is around 0.60 mm) lead to this result.

Tables 5, 6 and 7 present the results for the conditions *straight*, *ema-tongue* and *tm*, respectively, in the DNN setup. First of all, it becomes evident that DNN based modeling outperforms the HMM modeling even with a relatively small amount of data. The reason might by that the decision tree modeling is not able to capture some important correlation inside the data. Indeed, a decision tree clustering the space assumes that there is no connection between the different clusters. Therefore, we do not advise to use the default HMM configuration as the flexibility of the DNN will increase its accuracy if more data is added.

Therefore, it is more surprising that the configuration *ema-tongue* in the DNN setup is significantly worse than the other setup to predict the trajectories of the tongue coils. Our assumption is that the DNN doesn't have enough data to capture the bio-mechanical constraints of the tongue. This seems to be confirmed by the fact that using the parameters of the tongue model in the *tm* setup, does produce results comparable to the *ema* setup.

Comparing the distribution of distances across the setups and the phone classes (Figure 2), we find that only the tongue tip (coil T1) is more volatile and degrades the results. It is pos-

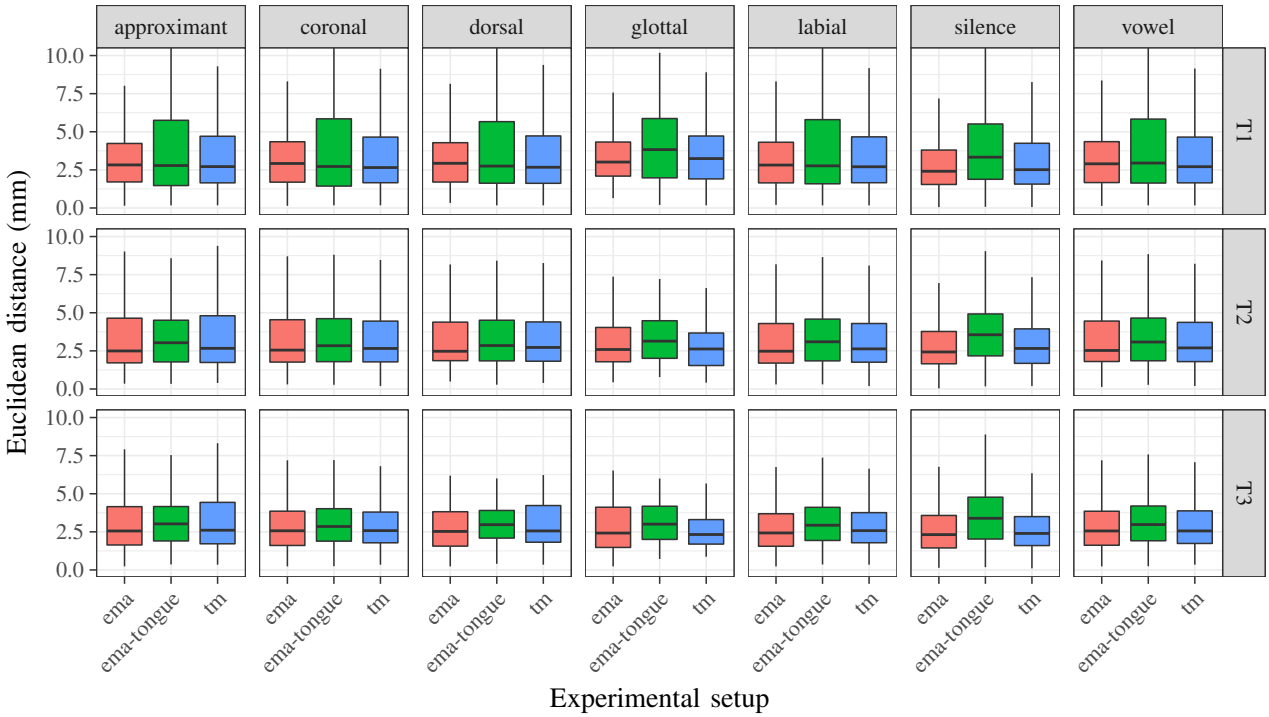

Figure 2: *Distributions of Euclidean distances between observed and predicted tongue EMA coil positions by the DNN modeling for each experimental TTS setup ("ema": EMA synthesis, "ema-tongue": EMA synthesis restricted to tongue coils, and "tm": synthesis of tongue model parameter), split by phone class and tongue EMA coil.*

Table 4: *HMM synthesis evaluation results for* tm *(synthesis of tongue model parameter).*

id	mean	std. dev.	conf. int.
eucdist T3 (mm)	2.61	1.61	9.43×10^{-3}
eucdist T2 (mm)	2.80	1.74	0.01
eucdist T1 (mm)	2.91	1.85	0.01

Table 5: straight *DNN synthesis evaluation results for* straight *(acoustic synthesis).*

id	mean	std. dev.	conf. int.
rms F_0 (cent)	153.62	67.86	10.91
rms F_0 (Hz)	8.57	5.06	0.81
vuv (%)	11.38	3.70	0.60
mcdist (dB)	2.13	0.20	0.03

sible that varying the smoothness term in the tongue model may improve this.

4. Conclusion

In this study, we have presented an objective comparison between HMM and DNN based modeling to synthesize acoustic speech and synchronized animation of a full 3D model of the tongue surface. First, we demonstrated a conventional, fused multimodal approach, then separated the two modalities while ensuring that the objective evaluation measures remained comparable in both modeling. Finally, we compared the results obtained by the HMM modeling to those obtained from the

Table 6: *DNN synthesis evaluation results for* ema-tongue *(EMA synthesis restricted to tongue coils).*

id	mean	std. dev.	conf. int.
eucdist T3 (mm)	3.84	2.08	0.01
eucdist T2 (mm)	3.97	2.07	0.01
eucdist T1 (mm)	3.75	2.41	0.01

Table 7: *DNN synthesis evaluation results for* tm *(synthesis of tongue model parameter).*

id	mean	std. dev.	conf. int.
eucdist T3 (mm)	2.07	1.35	7.92×10^{-3}
eucdist T2 (mm)	2.33	1.48	8.66×10^{-3}
eucdist T1 (mm)	2.22	1.49	8.74×10^{-3}

DNN approach of the acoustic and the tongue parameters. This demonstrates that even with a relatively small amount of data, the DNN approach already outperforms the HMM based modeling.

A number of synthesized utterances from the test set for both approaches are provided in the form of example videos in the multimedia supplement to this paper; no additional smoothing has been applied.

In future work, we plan to assess the impact on perceived naturalness by integrating the tongue model into a realistic talking avatar [e.g., 23, 24]. Regarding the tongue model integration, we plan to explore speaker adaptation using volumetric data, such as the MRI subset of the *mngu0* corpus [25].

5. References

[1] O. Engwall, "Evaluation of a system for concatenative articulatory visual speech synthesis," in *International Conference on Spoken Language Processing (ICSLP)*, Sep. 2002, pp. 665–668. [Online]. Available: http://www.isca-speech.org/archive/icslp_2002/i02_0665.html

[2] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis – determination, adjustment, evaluation," *Speech Communication*, vol. 44, no. 1-4, pp. 141–154, Oct. 2004.

[3] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, May 1987.

[4] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[5] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834–846, Oct. 2010.

[6] ——, "HMM-based text-to-articulatory-movement prediction and analysis of critical articulators," in *Interspeech*, Sep. 2010, pp. 2194–2197. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2010/i10_2194.html

[7] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Interspeech*, Aug. 2011, pp. 1505–1508. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_1505.html

[8] W. Katz, T. F. Campbell, J. Wang, E. Farrar, J. C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "OptiSpeech: a real-time, 3D visual feedback system for speech training." in *Interspeech*, Sep. 2014, pp. 1174–1178. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_1174.html

[9] 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 *Articulated Motion and Deformable Objects*. Springer, 2008, pp. 132–143.

[10] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, no. 2-3, pp. 303–329, Oct. 2003.

[11] K. James, A. Hewer, I. Steiner, and S. Wuhrer, "A real-time framework for visual feedback of articulatory data using statistical shape models," in *Interspeech*, Sep. 2016, pp. 1569–1570. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/2019.html

[12] I. Steiner, S. Le Maguer, and A. Hewer, "Synthesis of tongue motion and acoustics from text using a multimodal articulatory database," *IEEE Transactions on Audio, Speech, and Language Processing*, submitted, under revision. [Online]. Available: https://arxiv.org/abs/1612.09352

[13] A. Hewer, S. Wuhrer, I. Steiner, and K. Richmond, "A multilinear tongue model derived from speech related MRI data of the human vocal tract," *arXiv preprint arXiv:1612.05005*, 2016, submitted. [Online]. Available: https://arxiv.org/abs/1612.05005

[14] K. Richmond and S. Renals, "Ultrax: An animated midsagittal vocal tract display for speech therapy," in *Interspeech*, Sep. 2012, pp. 74–77. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2012/i12_0074.html

[15] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7962–7966.

[16] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[17] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Interspeech*, Sep. 2005. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_0093.html

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptative algorithm for mel-cepstral analysis of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Mar. 1992, pp. 137–140.

[20] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2000, pp. 1315–1318.

[21] A. Baker, "A biomechanical tongue model for speech production based on MRI live speaker data," 2011. [Online]. Available: http://www.adambaker.org/qmu.php

[22] HTS Working Group, *HTS Document: List of modifications made in HTS (for version 2.3)*, Dec. 2015. [Online]. Available: http://hts.sp.nitech.ac.jp/archives/2.3/HTS_Document.pdf

[23] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, Jul. 2012.

[24] D. Schabus, M. Pucher, and G. Hofer, "Joint audiovisual hidden semi-Markov model-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 336–347, Apr. 2014.

[25] I. Steiner, K. Richmond, I. Marshall, and C. D. Gray, "The magnetic resonance imaging subset of the mngu0 articulatory corpus," *Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL106–EL111, Feb. 2012.