Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis

Benjamin Picart, Thomas Drugman, Thierry Dutoit

TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Belgium
{benjamin.picart,thomas.drugman,thierry.dutoit}@umons.ac.be

10.21437/Interspeech.2011-32

Abstract

This paper focuses on the implementation of a continuous control of the degree of articulation (hypo/hyperarticulation) in the framework of HMM-based speech synthesis. The adaptation of a neutral speech synthesizer to generate hypo and hyperarticulated speech using a limited amount of speech data is first studied. This is done using inter-speaker voice adaptation techniques, applied here to intra-speaker voice adaptation. The implementation of a continuous control of the degree of articulation is then proposed in a second step. Finally, a subjective evaluation shows that good quality neutral/hypo/hyperarticulated speech, and also any intermediate, interpolated or extrapolated articulation degrees, can be obtained from an HMM-based speech synthesizer.

Index Terms: Speech Synthesis, HTS, Expressive Speech, Speaking Style Adaptation, Voice Quality

1. Introduction

The "H and H" theory [1] proposes two degrees of articulation of speech: hyperarticulated speech, for which speech clarity tends to be maximized, and hypoarticulated speech, where the speech signal is produced with minimal efforts. Therefore the degree of articulation provides information on the motivation/personality of the speaker vs the listeners [2]. Speakers can adopt a speaking style that allows them to be understood more easily in difficult communication situations. The degree of articulation is characterized by modifications of the phonetic context, of the speech rate and of the spectral dynamics (vocal tract rate of change). The common measure of the degree of articulation consists in defining formant targets for each phone, taking coarticulation into account, and studying differences between real observations and targets vs the speech rate. Since defining formant targets is not an easy task, Beller proposed in [2] a statistical measure of the degree of articulation by studying the joint evolution of the vocalic triangle area and the speech rate.

This paper is in line with our previous work on expressive speech synthesis [3]. We here focus on the synthesis of different speaking styles, with a varying degree of articulation: neutral speech, hypoarticulated (or casual) and hyperarticulated (or clear) speech. "Hyperarticulated speech" refers to the situation of a teacher/speaker talking in front of a large audience (important articulation efforts have to be made to be understood by everybody). "Hypoarticulated speech" refers to the situation of a person talking in a narrow environment or very close to someone (few articulation efforts have to be made to be understood). It is worth noting that these three modes of expressivity are neutral on the emotional point of view, but can vary amongst speakers, as reported in [2]. The influence of emotion on the articulation degree has been studied in [4] [5] and is out of the scope of this work.

In our previous work on the subject [3], an HMM-based speech synthesizer was built for each degree of articulation (neutral, hypo and hyper) using a large database for each degree of articulation. In this paper, we introduce the adaptation of a neutral synthesizer to generate hypo and hyperarticulated speech, using a limited amount of speech data. This is done using voice adaptation techniques in the spirit of [6] [7], but applied here to intra-speaker voice adaptation [8] - [11]. In particular, we study the efficiency of speaking style adaptation as a function of the size of the adaptation database. We then test the implementation of a continuous degree of articulation tuner, which is manually adjustable by the user to obtain not only neutral/hypo/hyperarticulated speech, but also any intermediate, interpolated or extrapolated articulation degrees, in a continuous way. Starting from an existing standard neutral voice with no hypo/hyperarticulated recordings available, the ultimate goal of our research is to allow for a continuous control of its articulation degree.

Hypo/hyperarticulated speech synthesis has many applications: expressive voice conversion (e.g. for embedded systems and video games), "reading speed" control for visually impaired people (i.e. fast speech synthesizers, more easily produced using hypoarticulation), ...

After a brief description of the contents of our database in Section 2, this paper is divided into two main parts. The first contribution addresses, in Section 3, the problem of speaking style adaptation (i.e. how to use a limited amount of sentences to adapt the model). The second contribution addresses, in Section 4, the implementation of a continuous control of the degree of articulation in the HMM-based speech synthesis system HTS [12]. Finally Section 5 concludes the paper.

2. Database with various Degrees of Articulation

For the purpose of our research, a new French database was recorded in [3] by a professional male speaker, aged 25 and native French (Belgium) speaking. The database contains three separate sets, each set corresponding to one degree of articulation (neutral, hypo and hyperarticulated). For each set, the speaker was asked to pronounce the same 1359 phonetically balanced sentences (around 75, 50 and 100 minutes of neutral, hypo and hyperarticulated speech respectively), as neutrally as possible from the emotional point of view. A headset was provided to the speaker for both hypo and hyperarticulated recordings, in order to induce him to speak naturally while modifying his articulation degree (see [3] for details on how this was induced).

Copyright © 2011 ISCA

3. Speaking Style Adaptation

We first study the adaptation of the neutral synthesizer trained in [3], to generate hypo and hyperarticulated speech with a limited amount of hypo and hyperarticulated speech data. Speaker adaptation can be performed by adapting an average voice model to a specific target speaker [6] [7]. The average voice model is computed once for all on a database containing many different speakers. In this work, we consider that our average voice model is the standard neutral HMM model. We use the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [13] [14] in the framework of the Hidden Semi Markov Model (HSMM) [15] to produce hypo/hyperarticulated speech. The linearly transformed models are further updated using MAP adaptation [7].

3.1. Method

An HMM-based speech synthesizer [16] was built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [12]. 1220 neutral sentences sampled at 16 kHz were used for the training, leaving around 10% of the database for synthesis. For the filter, we extracted the traditional Mel Generalized Cepstral (MGC) coefficients (with $\alpha = 0.42$, $\gamma = 0$ and order of MGC analysis = 24). For the excitation, we used the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in [17], since it was shown to significantly improve the naturalness of the delivered speech. More precisely, both deterministic and stochastic components of the DSM were estimated from the training dataset for each degree of articulation. Note that it was shown in [18] that only 1000 voiced frames are sufficient for a reliable estimation of these components. In this study, we used 75-dimensional MGC parameters (including Δ and Δ^2). Moreover, each covariance matrix of the state output and state duration distributions were diagonal.

For each degree of articulation, this neutral HMM-based speech synthesizer was adapted using CMLLR, with hypo/hyperarticulated speech data to produce a hypo/hyperarticulated HMM synthesizer. For each of the following evaluations, the full data models are the models trained on the entire training sets (1220 sentences, respectively neutral, hypo and hyperarticulated), and the adapted models are the models adapted from the neutral full data model, using from 5 to 1220 sentences of the hypo/hyperarticulated training sets. Evaluations are performed on the test set, composed of sentences which were neither part of the training set nor of the adaptation set.

3.2. Objective Evaluation

The goal of the objective evaluation is to assess the quality of the synthesized speech when the number of adaptation sentences increases from 5 to 1220. All the measures presented here below are computed for all the vowels of the test set.

Like in [6], three objective measures are computed between the adapted and the full data models, as illustrated in Figure 1: the average mel-cepstral distortion (expressed in decibel), the root-mean-square (RMS) error of log F0 (expressed in cent), the RMS error of vowel durations (expressed in terms of number of frame). Note that since log F0 is not observed in unvoiced regions, the RMS error is computed for regions where both the adapted and the full data models are voiced. Cent is a logarithmic unit used for musical intervals (100 cents correspond to a semitone). Note that, in this objective evaluation, in order to have an alignment between phones generated from the adapted models and phones obtained from the full data models, the HMM-based speech synthesizer was forced to use the original phone durations (as pronounced by the speaker) for the computation of the average mel-cepstral distortion and the RMS error of log F0. These measures reflect differences regarding three complementary aspects of speech.

From Figure 1, it is observed that adapted hyperarticulated speech is always further away from the full data model than adapted hypoarticulated speech. As expected, mel-cepstral distortion, RMS error of log F0 and RMS error of vowel duration decrease when the number of adaptation sentences increases. It is particularly strong for the mel-cepstral distortion. However, a bit less than 100 sentences, i.e. around 3 (7) minutes of hypo (hyper) articulated speech, are sufficient to adapt F0 and phone duration with a good quality, while around 200 sentences, i.e. around 7 (13) minutes of hypo (hyper) articulated speech, are needed to adapt cepstra with a good quality. Indeed 1 dB is usually accepted as the difference limen for spectral transparency [19]. Note that for a same amount of sentences, differences between durations come from different speech rates in hypo and hyperarticulated speech [3]. For the inter-speaker adaptation [6], average voice models were trained and adapted using from 5 to 450 sentences. Despite some differences in the training process and in the number of training and adaptation data, the same kind of tendency can be observed here for intra-speaker adaptation.

3.3. Subjective Evaluation

In order to confirm the objective evaluation conclusions, we performed a Comparison Category Rating (CCR) evaluation. For this evaluation, listeners were asked to listen to two sentences: A, the sentence synthesized by the full data model; B, the sentence synthesized by the adapted model using 10, 20, 50, 100 or 1220 sentences. The CCR values range on a gradual scale varying from 1 (meaning that A and B are very dissimilar) to 5 (meaning the opposite). A score of 3 is given if both versions are found to be slightly similar. Listeners were asked to score the overall speech quality (voice characteristics and prosodic features) of B compared to A. The higher the CCR score, the more efficient the adaptation process. Note that here the HMM-based speech synthesizer generates cepstra, F0 and duration, conversely to the objective evaluation explained in Section 3.2.

The test consisted of 30 pairwise comparisons. For each degree of articulation, 5 sentences were randomly chosen from the test set. During the test, listeners were allowed to listen to each pair of sentences as many times as wanted, in the order they preferred. However they were not allowed to come back to previous sentences after validating their decision. 26 people, mainly naive listeners, participated to this evaluation. The mean CCR score, together with its 95% confidence interval for each articulation degree, is shown in Figure 2. This graph confirms the objective evaluation results. Adapted hypoarticulated speech is better rendered by HTS than adapted hyperarticulated speech. The quality is already good when adapting the model with 100 sentences, but the more adaptation sentences, the better the quality independently of the degree of articulation.

4. Interpolation/Extrapolation of the Degree of Articulation

The second contribution of this work is to implement and test a continuous control of the degree of articulation in HMM-based speech synthesis, in order to smoothly and continuously change

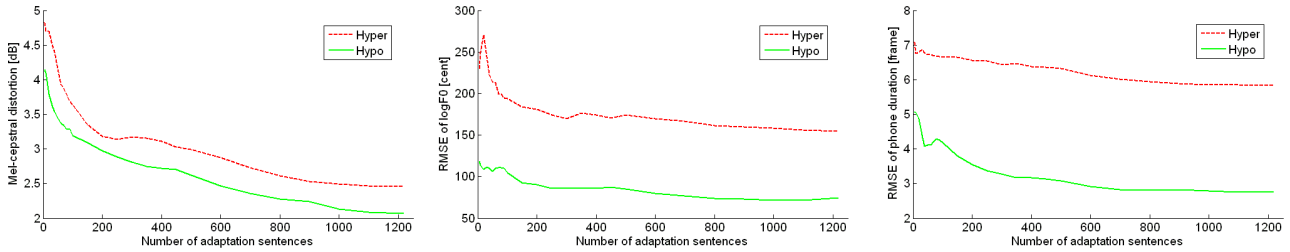

Figure 1: Objective measures computed between the adapted and the full data models: (left) average mel-cepstral distortion [dB]; (middle) RMS error of log F0 [cent]; (right) RMS error of vowel durations [number of frame].

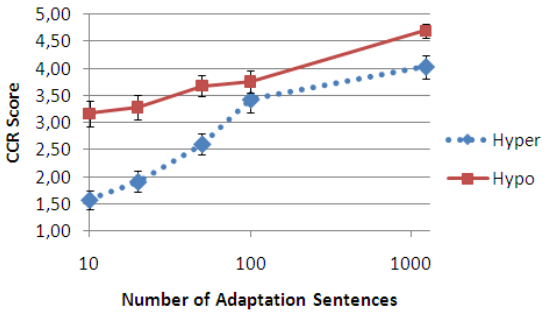

Figure 2: Subjective evaluation - Effect of the number of adaptation sentences, on CCR scores.

the degree of articulation of the neutral voice towards and beyond our adapted hypo or hyperarticulated voices.

Thanks to the parametric representation used in HMM-based speech synthesis, interpolation between the speaking styles is possible. Speaker interpolation is performed in [20] by interpolating HMM parameters among some representative speakers HMM sets. They assume that each HMM state has a single Gaussian output distribution, reducing the problem to the interpolation amongst N Gaussian distributions. Yamagishi proposed three methods for modeling and interpolating between speaking styles [8]: style dependent modeling and style mixed modeling [22]; model interpolation technique [20]; MLLR-based model adaptation technique [21]. In this latter study, speaking styles means emotions, while it means degree of articulation in our case. Dialect interpolation has been performed in [23] using dialect-dependent and dialect-independent modelings. [24] suggests factor analyzed voice models for creating various voice characteristics in the HMM-based speech synthesis.

4.1. Method

The training and adaptation of the HMM-based speech synthesizer is performed in the same way as in Section 3. In order to obtain the best possible speech quality with the adaptation process, the entire training set (1220 sentences) was used. The continuous control of the degree of articulation is achieved by linearly interpolating/extrapolating the mean and the diagonal covariance matrices of each state output and state duration probability density functions (mel-cepstrum, log F0 and duration distributions). Since no reference speech data is available to evaluate objectively the quality of the interpolation/extrapolation, a subjective test is conducted.

4.2. Subjective Evaluation

For this evaluation, listeners were asked to listen to four sentences: the three reference sentences A (hypo), B (neutral) and C (hyper) synthesized by the full data models; the test sentence X, which could be either interpolated between A and B or B and C, or extrapolated beyond A or C. Then they were given a discrete scale, ranging from -1.5 to 1.5 by a 0.25 step. A, B and C were placed at -1, 0 and 1 respectively. Finally, they were asked to tell where X should be located on that scale, X being different from A, B or C. They were also asked to score the overall speech quality of X versus B (the neutral synthesis), leaving aside the difference in articulation between X and B. For this, we used a Comparative Mean Opinion Score (CMOS) test in order to assess the quality of the interpolated/extrapolated speech synthesis. CMOS values range on a gradual scale varying from -3 (meaning that X is much worse than B) to +3 (meaning the opposite). A score of 0 is given if the quality of both versions is found to be equivalent.

The test consisted of 10 quadruplets. For each degree of articulation, 5 sentences were randomly chosen from the test set. We used the same listening conditions as in Section 3.3. 34 people, mainly naive listeners, participated to this evaluation. Table 1 displays the evolution of the average perceived interpolation/extrapolation ratio, as a function of the actual ratio which is applied. 95% confidence intervals are also indicated. We clearly see that the perceived degree of articulation corresponds quite well to the reference degree of articulation. However, due to the fact that the user was not allowed to select reference (-1, 0, 1) or extreme (lower than -1.5, higher than 1.5) values, we may have introduced a small bias in the assessment of the perceived degree of articulation.

The averaged CMOS scores, corresponding to the perceived synthesis quality, together with their 95% confidence intervals, are shown in Figure 3. From this figure, interpolated hyperarticulation seems to have about the same quality as neutral speech, while a slight degradation is observed for all other degrees of articulation. Notice that since the degree of articulation of X and B could be different, it is hard to compare speech quality alone, explaining the size of the large 95% confidence intervals.

5. Conclusions

We have implemented a continuous control of the degree of articulation (hypo/hyperarticulation) in the framework of HMM-based speech synthesis. In a first step, we performed the adaptation of a neutral synthesizer to generate hypo and hyperarticulated speech with a limited amount of speech data. An objective evaluation showed that, for intra-speaker adaptation, a bit less than 100 sentences, i.e. around 3 (7) minutes of hypo

Table 1: *Subjective evaluation - Perceived interpolation/extrapolation ratio, together with its 95% confidence interval, versus the actual interpolation/extrapolation ratio.*

	Hypoarticulation		Hyperarticulation	
	Actual	Perceived	Actual	Perceived
Degree of Articulation	0.25	0.50 ± 0.13	- 0.25	- 0.40 ± 0.18
	0.5	0.54 ± 0.13	- 0.5	- 0.55 ± 0.11
	0.75	0.77 ± 0.10	- 0.75	- 0.71 ± 0.12
	1.25	1.07 ± 0.08	- 1.25	- 1.05 ± 0.11
	1.5	1.09 ± 0.13	- 1.5	- 1.23 ± 0.10

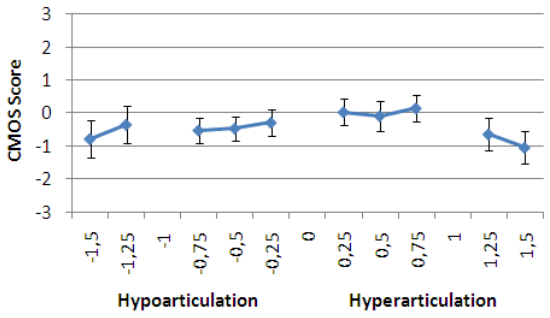

Figure 3: Subjective evaluation - Perceived synthesis quality of the test sentence X vs the neutral sentence B.

(hyper) articulated speech, are sufficient to adapt F0 and phone duration, while around 200 sentences, i.e. around 7 (13) minutes of hypo (hyper) articulated speech, are needed to adapt cepstra with a good quality, which is quite equivalent to the tendency observed for inter-speaker adaptation. These results were confirmed by a subjective test. In a second step, the implementation of a continuous control of the articulation degree was proposed. Subjective evaluation showed that good quality neutral/hypo/hyperarticulated speech, but also any intermediate, interpolated or extrapolated articulation degrees, can be obtained from an HMM-based speech synthesizer. This work is a first step towards an automatic continuous control of the degree of articulation on an existing standard neutral voice, with no hypo/hyperarticulated recordings available.

Audio examples for speaking style adaptation and for interpolation/extrapolation of the degree of articulation are available online via http://tcts.fpms.ac.be/~picart/.

6. Acknowledgements

Benjamin Picart is supported by the “Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (FRIA). Thomas Drugman is supported by the “Fonds National de la Recherche Scientifique” (FNRS).

7. References

[1] B. Lindblom, Economy of Speech Gestures, vol. The Production of Speech, Spinger-Verlag, New-York, 1983.

[2] G. Beller, *Analyse et Modèle Génératif de l’Expressivité - Application à la Parole et à l’Interprétation Musicale*, PhD Thesis (in French), Universit Paris VI - Pierre et Marie Curie, IRCAM, 2009.

[3] B. Picart, T. Drugman, T. Dutoit, *Analysis and Synthesis of Hypo and Hyperarticulated Speech*, Proc. Speech Synthesis Workshop 7 (SSW7), Kyoto, Japan, 2010.

[4] G. Beller, *Influence de l’expressivité sur le degré d’articulation*, RJCP, France, 2007.

[5] G. Beller, N. Obin, X. Rodet, *Articulation Degree as a Prosodic Dimension of Expressive Speech*, Fourth International Conference on Speech Prosody, Campinas, Brazil, 2008.

[6] J. Yamagishi, T. Kobayashi, *Average-Voice-based Speech Synthesis using HSMM-based Speaker Adaptation and Adaptive Training*, IEICE Trans. Information and Systems, E90-D, no. 2, pp. 533-543, 2007.

[7] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, *A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis*, IEEE Audio, Speech, & Language Processing, vol. 17, no. 6, pp. 1208-1230, August 2009.

[8] J. Yamagishi, T. Masuko, T. Kobayashi, *HMM-based expressive speech synthesis – Towards TTS with arbitrary speaking styles and emotions*, Proc. of Special Workshop in Maui (SWIM), 2004.

[9] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, *HMM-based speech synthesis with various speaking styles using model interpolation and adaptation*, IEICE Technical Report, vol. 103, no. 264, pp. 37-42, 2003.

[10] T. Nose, J. Yamagishi, T. Masuko, T. Kobayashi, *A style control technique for HMM-based expressive speech synthesis*, IEICE Trans. on Information and Systems, vol. 90, no. 9, pp. 1406-1413, 2007.

[11] T. Nose, M. Tachibana, T. Kobayashi, *HMM-Based Style Control for Expressive Speech Synthesis with Arbitrary Speaker’s Voice Using Model Adaptation*, IEICE Transactions on Information and Systems, vol. 92, no. 3, pp. 489-497, 2009.

[12] [Online] HMM-based Speech Synthesis System (HTS) website : http://hts.sp.nitech.ac.jp/

[13] V. Digalakis, D. Rtischev, L. Neumeyer, *Speaker adaptation using constrained reestimation of Gaussian mixtures*, IEEE Trans. Speech Audio Process., vol. 3, no. 5, pp. 357-366, 1995.

[14] M. Gales, *Maximum likelihood linear transformations for HMM-based speech recognition*, Comput. Speech Lang., vol. 12, no. 2, pp. 75-98, 1998.

[15] J. Ferguson, *Variable Duration Models for Speech*, in Proc. Symp. on the Application of Hidden Markov Models to Text and Speech, pp. 143-179, 1980.

[16] H. Zen, K. Tokuda, A. W. Black, *Statistical parametric speech synthesis*, Speech Commun., vol. 51, no. 11, pp. 1039-1064, 2009.

[17] T. Drugman, G. Wilfart, T. Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Proc. Interspeech, Brighton, U.K., 2009.

[18] T. Drugman, T. Dutoit, *On the Potential of Glottal Signatures for Speaker Recognition*, Proc. Interspeech, Makuhari, Japan, 2010.

[19] K. K. Paliwal, B. S. Atal, *Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame*, IEEE Trans. Speech Audio Process., vol. 1, no. 1, pp. 3-14, 1993.

[20] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, T. Kitamura, *Speaker interpolation for HMM-based speech synthesis system*, Journal of the Acoustic Society of Japan (E), vol. 21, no. 4, pp.199-206, 2000.

[21] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, *Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR*, IEEE ICASSP, pp. 805-808, Salt Lake City, USA, 2001.

[22] J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, *Modeling of various speaking styles and emotions for HMM-based speech synthesis*, in Proc. Eurospeech, pp. 2461-2464, Switzerland, 2003.

[23] M. Puchera, D. Schabus, J. Yamagishi, F. Neubarth, V. Strom, *Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis*, Speech Commun., vol. 52, no. 2, pp. 164-179, 2010.

[24] K. Kazumi, Y. Nankaku, K. Tokuda, *Factor analyzed voice models for HMM-based speech synthesis*, IEEE ICASSP, pp. 4234-4237, Dallas, Texas, 2010.