# Perceptual Effects 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

**Abstract.** This paper focuses on the understanding of the effects leading to high-quality HMM-based speech synthesis with various degrees of articulation. The adaptation of a neutral speech synthesizer to generate hypo and hyperarticulated speech is first performed. The impact of cepstral adaptation, of prosody, of phonetic transcription as well as the adaptation technique on the perceived degree of articulation is studied. For this, a subjective evaluation is conducted. It is shown that high-quality hypo and hyperarticulated speech synthesis requires the use of an efficient adaptation such as CMLLR. Moreover, in addition to prosody adaptation, the importance of cepstrum adaptation as well as the use of a Natural Language Processor able to generate realistic hypo and hyperarticulated phonetic transcriptions is assessed.

## 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 (i.e. shape formed by vowels in the $F1$ - $F2$ space) and the speech rate.

We focus on the synthesis of different speaking styles, with a varying degree of articulation: neutral, hypoarticulated (or casual) and hyperarticulated (or clear) speech. "Hyperarticulated speech" refers to the situation of a 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 [3] and is out of the scope of this work.

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), ...

This paper is in line with our previous works on expressive speech synthesis. The analysis and synthesis of hypo and hyperarticulated speech, in the framework of Hidden Markov Models (HMMs), has been performed in [4]. Significant differences between the three degrees of articulation were shown, both on acoustic and phonetic aspects. We then studied the efficiency of speaking style adaptation as a function of the size of the adaptation database [5]. Speaker adaptation [6] is a technique to transform a source speaker's voice into a target speaker's voice, by adapting the source HMM-based model (which is trained using the source speech data) with a limited amount of target speech data. The same idea lies for speaking style adaptation [7] [8]. We were therefore able to produce neutral/hypo/hyperarticulated speech directly from the neutral synthesizer. We finally implemented a continuous control (tuner) of the degree of articulation on the neutral synthesizer [5]. This tuner was 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.

This paper focuses on a deeper understanding of the phenomena responsible in the perception of the degree of articulation. This perceptual study is necessary as a preliminary step towards performing a speaker-independent control of the degree of articulation. Indeed the articulation degree induces modifications in the cepstrum, pitch, phone duration and phonetic transcription. In this work, these modifications are analyzed and quantified in comparison with a baseline, in which a straightforward, phone-independent constant ratio is applied to the pitch and phone durations of the neutral synthesizer in order to get as close as possible to real hypo/hyperarticulated speech.

After a brief description of the contents of our database in Section 2, the implementation of our synthesizers in the HMM-based speech synthesis system HTS [9] is detailed in Section 3. Results with regard to effects influencing the perception of the degree of articulation are given in Section 4. Finally Section 5 concludes the paper.

## 2 Database

For the purpose of our research, a new French database was recorded in [4] 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 [4].

## 3 Conception of an HMM-based Speech Synthesizer

An HMM-based speech synthesizer [10] was built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [9]. 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 [11], since it was shown to significantly improve the naturalness of the delivered speech. More precisely, both deterministic and stochastic components of DSM were estimated on the training dataset for each degree of articulation. In this study, we used 75-dimensional MGC parameters (including $\Delta$ and $\Delta^2$). 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 the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [12] [13] in the framework of the Hidden Semi Markov Model (HSMM) [14], with hypo/hyperarticulated speech data to produce a hypo/hyperarticulated speech synthesizer. HSMM is an HMM having explicit state duration distributions (advantage during the adaptation process of phone duration). CMLLR is a feature adaptation technique which estimates a set of linear transformations for the features so that each state in the HMM system is more likely to generate the adaptation data. The linearly transformed models are further updated using a Maximum A Posteriori (MAP) adaptation [6].

In the following, the full data models refer to 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 hypo/hyperarticulated speech data. We showed in [5] that good quality adapted models can be obtained when adapting the neutral full data model with around 100-200 hypo/hyperarticulated sentences. On the other hand, the more adaptation sentences, the better the quality independently of the degree of articulation. This is why we chose in this work to adapt the neutral full data model using the entire hypo/hyperarticulated training sets. This will also allow us to remove from our results the amount of adaptation data from the possible perceptual effects (as it is studied in [5]).

Based on the full data models and on the adapted models, four synthesizers are created: one for each effect to be analyzed, as summarized in Table 1. For example, *Case 1* is our baseline system and corresponds to the neutral full data

model, where a straightforward phone-independent constant ratio is applied to decrease/increase pitch and phone durations to sound like hypo/hyperarticulated speech. This ratio is computed once for all over the hypo/hyperarticulated databases (see Section 2) by adapting the mean values of the pitch and phone duration from the neutral style. The phonetic transcription is manually adjusted to fit the real hypo/hyperarticulated transcription.

**Table 1.** *Conception of four different synthesizers, each of them focusing on an effect influencing the degree of articulation.*

| | Full Data Model (Neutral) | | | | Adapted Model (Hypo/Hyper) | | | |
|---|---|---|---|---|---|---|---|---|
| | Cepstrum | Pitch | Duration | Phon. Transcr. | Cepstrum | Pitch | Duration | Phon. Transcr. |
| Case 1 | X | Ratio | Ratio | | | | | X |
| Case 2 | X | | | | | X | X | X |
| Case 3 | | | | X | X | X | X | |
| Case 4 | | | | | X | X | X | X |

## 4 Experiments

In order to evaluate the influence of each factor explained in Section 3, a subjective test is conducted. For this evaluation, listeners were asked to listen to three sentences: the two reference sentences A (neutral) and B (hypo/hyper) synthesized by the full data models; the test sentence X synthesized by one of the four synthesizers described in Table 1 (randomly chosen), which could be either hypo or hyperarticulated depending on the articulation of B. Then participants were given a continuous scale, ranging from -0.25 to 1.25. A and B were placed at 0 and 1 respectively. Given this, they were asked to tell where X should be located on that scale. Evaluation was performed on the test set, composed of sentences which were neither part of the training set nor of the adaptation set.

The test consisted of 20 triplets. For each degree of articulation, 10 sentences were randomly chosen from the test set. During the test, listeners were allowed to listen to each triplet 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. 24 people, mainly naive listeners, participated to this evaluation. The mean Perceived Degree of Articulation (PDA) scores, together with their 95% confidence intervals are shown in Figure 1. The closer to 1 the PDA scores, the better the synthesizer as it leads to an efficient rendering of the intended degree of articulation.

From this figure, we clearly see the advantage of using an HMM to generate prosody (pitch and phone duration) instead of applying a straightforward phone-independent constant ratio to the neutral synthesizer prosody, in order to get as close as possible to real hypo/hyperarticulated speech (*Case 1* vs *Cases 2*,

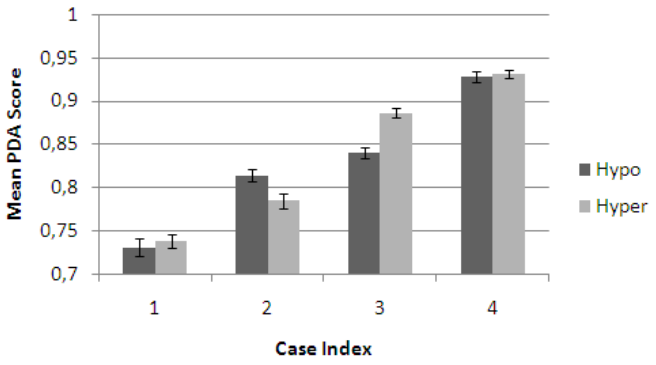


**Fig. 1.** Subjective evaluation - Mean PDA scores with their 95% confidence intervals (CI) for each degree of articulation.

*3*, *4*). The effects of cepstrum adaptation (*Case 2* vs *Case 4*) and phonetic adaptation (*Case 3* vs *Case 4*) are also highlighted. It can be noted that adapting the cepstrum has a higher impact on the rendering of the articulation degree than adapting the phonetic transcription. Moreover, it is also noted that this conclusion is particularly true for hyperarticulated speech, while the difference is less marked for hypoarticulation. When analyzing *Case 3*, it is observed that a lack of appropriate phonetic transcription is more severe for hypoarticulated speech. Indeed, we have shown in [4] that hypoarticulated speech is characterized by a high number of deletions, which is more important than the effect of phone insertions for hyperarticulated speech. It is also noticed for *Case 2* that the influence of spectral features is more dominant for hyperarticulated speech. This might be explained by the fact that spectral changes (compared to the neutral style) induced by an hyperarticulation strategy are important to be modeled by the HMMs. Although significant spectral modifications are also present for hypoarticulated speech, it seems that their impact on the listener perception is marked to a lesser extent. Finally, it is noted that a high performance is achieved by the complete adaptation process (*Case 4* vs ideal value 1, which is the speech synthesized using the full data hypo/hyperarticulated models). This proves the efficiency of the degree of articulation CMLLR adaptation based on HMM.

## 5 Conclusions

In this paper, HMM proved its usefulness in modeling high-quality hypo and hyperarticulated speech. Indeed adaptation of cepstrum, pitch and phone duration from the neutral full data model outperforms the baseline, in which a straightforward phone-independent constant ratio is applied to pitch and phone durations to get as close as possible to real hypo/hyperarticulated speech. We also highlighted the fact that adapting prosody alone, without adapting cep-

strum highly degrades the rendering of the degree of articulation. The importance of having a Natural Language Processor able to create automatically realistic hypo/hyperarticulated transcriptions has been proven. Finally, the impact of cepstrum adaptation is more important than the effect of phonetic transcription.

Audio examples for each effect responsible for the perception of the degree of articulation are available online via http://tcts.fpms.ac.be/~picart/.

## Acknowledgments

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).

## 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. G. Beller, N. Obin, X. Rodet, *Articulation Degree as a Prosodic Dimension of Expressive Speech*, Fourth International Conference on Speech Prosody, Campinas, Brazil, 2008.
4. B. Picart, T. Drugman, T. Dutoit, *Analysis and Synthesis of Hypo and Hyperarticulated Speech*, Proc. Speech Synthesis Workshop 7 (SSW7), Kyoto, Japan, 2010.
5. B. Picart, T. Drugman, T. Dutoit, *Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis*, Proc. Interspeech, Firenze, Italy, 2011.
6. 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.
7. 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.
8. 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.
9. HMM-based Speech Synthesis System (HTS) website : http://hts.sp.nitech.ac.jp/
10. H. Zen, K. Tokuda, A. W. Black, *Statistical parametric speech synthesis*, Speech Commun., vol. 51, no. 11, pp. 1039-1064, 2009.
11. 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.
12. 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.
13. M. Gales, *Maximum likelihood linear transformations for HMM-based speech recognition*, Comput. Speech Lang., vol. 12, no. 2, pp. 75-98, 1998.
14. 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.