Assessing the Intelligibility and Quality of HMM-based Speech Synthesis with a Variable Degree of Articulation

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

Abstract

This paper focuses on the assessment of both the intelligibility and the quality of speech when using a variable degree of articulation (hypo/hyperarticulation) in the framework of HMM-based speech synthesis. Intelligibility is evaluated when the synthesizer is working in adverse conditions. The adaptation of a neutral speech synthesizer to generate hypo and hyperarticulated speech is first performed. Simulated noisy and reverberant conditions are then applied to the speech produced by the latter synthesizers. The intelligibility of the resulting speech is assessed by a Semantically Unpredictable Sentences (SUS) test. Results of this test quantify how the possibility of varying the degree of articulation improves the intelligibility of synthetic speech in various adverse conditions. In a second test, natural and synthetic speech quality is evaluated through an Absolute Category Rating (ACR) test. This test allows the assessment of hypo/hyperarticulated speech through various dimensions: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation.

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

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 allowing 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.

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), etc.

This paper is in line with our previous works on expressive speech synthesis [3] [4] [5]. 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 [6] [7] and is out of the scope of this work.

In our previous work on the topic [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. We then studied the efficiency of speaking style adaptation as a function of the size of the adaptation database [4]. Speaker adaptation [8] 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 [9] [10]. 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 [4]. 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. Finally, we conducted a perceptual evaluation in [5] in order to have a deeper understanding of the phenomena responsible in the perception of the degree of articulation. 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.

In this work, we evaluate the necessity of integrating a variable degree of articulation in a HMM-based speech synthesis system when this latter is embedded in adverse conditions. This situation happens very often in concrete applications: for example, GPS voice inside a moving car (additive noise), train/flight information in stations/halls (reverberation), etc. Does hypo/hyperarticulated speech provides better intelligibility performance in adverse environments? What are the advantages brought by hypo/hyperarticulated speech compared to the standard neutral speech? This work is designed to provide an answer to these two questions.

This paper is structured as follows. 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 [11] is detailed in Section 3. Speech intelligibility in both noisy and reverberant environments, and speech quality evalu-

ations are performed in Sections 4 and 5. These tests quantify the usefulness of integrating the degree of articulation within HMM-based speech synthesis. Finally Section 6 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).

3. Implementation of the Speech Synthesizers

An HMM-based speech synthesizer [12] was built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [11]. 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 α = 0.42, γ = 0 and order of MGC analysis = 24). For the excitation, we used the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in [13], 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 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 the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [14] [15] in the framework of the Hidden Semi Markov Model (HSMM) [16], with hypo/hyperarticulated speech data to produce a hypo/hyperarticulated speech synthesizer.

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 [4] 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.

Finally, a linear interpolation of the means and the covariance matrices of each state output and state duration probability density functions (mel-cepstrum, log F0 and duration distributions) is computed between the neutral full data model and the adapted models. For experiments in this work, we chose an interpolation ratio equal to 0.5, corresponding to a model right between the neutral full data model and, on the one hand the adapted hypoarticulated model, and on the other hand the adapted hyperarticulated model.

4. Semantically Unpredictable Sentences Test

In order to evaluate the intelligibility of a voice, the Semantically Unpredictable Sentences (SUS) test was performed on speech degraded alternatively by an additive or a convolutive noise. The advantage of such sentences is that they are unpredictable, meaning that the listeners cannot determine a word in the sentence by the meaning of the whole utterance or the context within the sentence.

4.1. Building the SUS Corpus

The same corpus as the one built in [17] was used in our experiments. This corpus is part of the ELRA package (ELRA-E0023). Basically, 288 semantically unpredictable sentences were generated following 4 syntactic structures:

- adverb det. Noun_1 Verb-t-pron. det. Noun_2 Adjective?
- determiner Noun_1 Adjective Verb determiner Noun_2.
- det. Noun_1 Verb_1 determiner Noun_2 qui (that) Verb_2.
- determiner Noun_1 Verb preposition determiner Noun_2.

Structure 3 originally proposed by [18] was not kept, because it only contained 3 target words (nouns, verbs or adjectives, here written with a capital initial letter) instead of 4 in the other structures. For more details about the generation of this corpus, the reader is referred to [17].

4.2. Procedure

Nineteen people, mainly naive listeners, participated to this evaluation. They were asked to listen to 40 SUS, randomly chosen from the SUS corpus built in the previous paragraph. The SUS were played one at a time. For each of them, listeners were asked to write down what they heard. During the test, listeners were allowed to listen to each SUS at most two times. They were of course not allowed to come back to previous sentences after validating their decision.

The SUS were synthesized using the five synthesizers described in Section 3: neutral (0), hypo (-1) and hyperarticulated (1), interpolated between neutral and hypo (-0.5), and interpolated between neutral and hyper (0.5).

For simulating the noisy environment, a car noise was added to the original speech waveform at two Signal-to-Noise Ratios (SNRs): -5dB and -15dB. The car noise signal was taken from the Noisex-92 database [19], and was added so as to control the overall SNR without silence removal. Since the spectral energy of the car noise is mainly concentrated in the low frequencies ($<$400Hz), the formant structure of speech was only poorly altered, and voices remained somehow understandable even for SNR values as low as -15dB.

When the speech signal $s(n)$ is produced in a reverberant environment, the observation $x(n)$ at the microphone is:

$$x(n) = h(n) * s(n), \quad (1)$$

where $h(n)$ is the L-tap Room Impulse Response (RIR) of the acoustic channel between the source and the microphone. RIRs are characterized by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. In order to produce reverberant speech, a room measuring 3x4x5 m

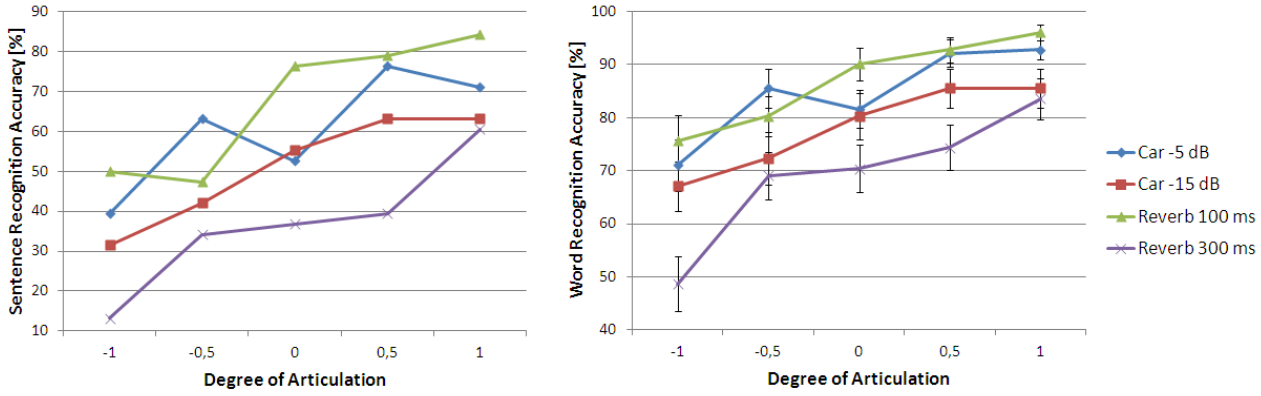

Figure 1: SUS Test - Mean sentence (left) and word (right) recognition accuracies [%].

with two levels of reverberation (T_{60} of 100 and 300ms) was simulated using the source-image method [20], and the simulated impulse responses convolved with original speech signals.

4.3. Results

The mean recognition accuracies at the sentence and word levels (for each degree of articulation, for each type and level of perturbation) are shown in Figure 1. The higher the score, the better the synthesizer intelligibility as it leads to higher sentence/word recognition accuracies. 95% confidence intervals are also displayed for information.

In order to cope with orthographic mistakes, these accuracies were manually annotated, by counting the number of erroneous phonemes for each sentence and word written by the listeners, in comparison with the correct sentence and word. A strong correlation is noted between the recognition accuracy at the sentence and word levels. For the computation of the results, a single erroneous phoneme inside the sentence leads to consider the sentence as wrong. The same idea was applied to the word recognition accuracy computation. Therefore, a sentence could be considered as wrong while some of its words could be considered as correct. Interestingly, accuracy increases with the degree of articulation, and decreases when the perturbation level rises. The worst adverse condition turns out to be the most severe reverberation. Finally, it is noted that the two values of hyperarticulated degree (0.5 and 1) lead to almost the same performance in a car noise, which implies that there is no need to increase the degree of articulation beyond 0.5 in such an environment. This conclusion however does not hold with a reverberant perturbation.

5. Absolute Category Rating Test

Finally, an Absolute Category Rating (ACR) test was conducted in order to assess the quality of speech. As in [17], the Mean Opinion Score (MOS) was complemented with six other categories: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation.

5.1. Procedure

Seventeen people, mainly naive listeners, participated to this evaluation. They were asked to listen to 18 meaningful sentences, randomly chosen amongst the held-out set of the database (used neither for training nor for adaptation). The sentences were played one at a time. For each of them, listeners were asked to rate according to the 7 aspects cited above (for the detailed questions list, see [17]). Listeners were given 7 continuous scales (one for each question to answer) ranging from 1 to 5. These scales were extended one point further on both sides (ranging therefore from 0 to 6) in order to prevent border effects. The sentences corresponded either to the original speech or to the synthesized speech with a variable degree of articulation (neutral, hypo/hyperarticulated). During the test, listeners were allowed to listen to each sentence as many times as wanted. However they were not allowed to come back to previous sentences after validating their decision.

5.2. Results

Results together with their 95% confidence intervals are shown in Figure 2. In all cases, original speech is preferred to synthetic speech. The MOS test shows that original neutral speech is preferred to hypo/hyperarticulated speech, while synthetic neutral and hyperarticulated speech are almost equivalent, leaving synthetic hypoarticulated speech slightly below. The comprehension test points out that neutral and hyperarticulated speech are clearly more understandable than hypoarticulated speech, both on the original and synthetic side. Differences of comprehension between original and synthesized speech are interestingly weak. The pleasantness test indicates a preference of the listeners for original neutral speech, followed by hyper and hypoarticulated speech, while all the types of synthetic speech are equivalently preferred. Despite the HMM modeling, the intonation and dynamics of the voice is well reproduced at synthesis, as illustrated with the non-monotony test. A major problem with HMM-based speech synthesis is the naturalness of the generated speech compared to the original speech. This is a known problem related in many studies. The naturalness test underlines again this conclusion. The fluidity test has an "inverse" tendency compared to other tests. Indeed hypoarticulated speech has a higher score than the others. This is due to the fact that hypoarticulated speech is characterized by a lower number of breaks and glottal stops, shorter phone durations and higher speech rate (as proven in [3]). All these effects lead to an impression of fluidity in speech, while the opposite tendency is observed in hyperarticulated speech. Finally, the pronunciation test correlates with the comprehension test in the sense that the more pronunciation problems are found, the harder the understandability of the message.

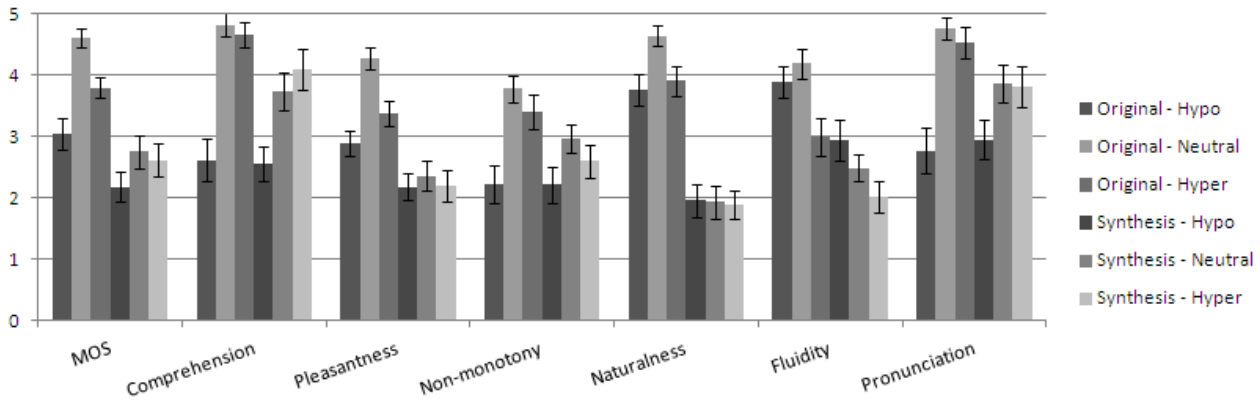

Figure 2: ACR Test - Mean score together with their 95% confidence intervals.

6. Conclusions

This paper focused on the evaluation of HMM-based speech synthesis integrating a variable degree of articulation. In a first step, the intelligibility of speech was assessed through a SUS test. In presence of a perturbation, this evaluation showed that hyperarticulated speech enhances the comprehension of synthetic speech. Moreover, a degree of articulation of 0.5 (instead of 1) is sufficient to improve the recognition of the message in a car noise. The same conclusion is drawn in reverberant environments, except that a degree of articulation of 1 is in this case necessary. In a second step, the quality of speech was assessed by an ACR test. This evaluation showed the gap in naturalness and pleasantness between original speech and synthetic speech. However, it is worth emphasizing that comprehension, non-monotony and pronunciation are well reproduced after the HMM modeling.

Audio examples related to our studies are available online at http://tcts.fpms.ac.be/~picart/.

7. 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 Walloon Region, SPORTIC project #1017095.

8. 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), pp. 270-275, Kyoto, Japan, 2010.

[4] B. Picart, T. Drugman, T. Dutoit, *Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis*, Proc. Interspeech, pp. 1797-1800, Firenze, Italy, 2011.

[5] B. Picart, T. Drugman, T. Dutoit, *Perceptual Effects of the Degree of Articulation in HMM-based Speech Synthesis*, Proc. NOLISP Workshop, pp. 177-182, Las Palmas, Gran Canaria, 2011.

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

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] P. Boula de Mareüil, C. d’Alessandro, A. Raake, G. Bailly, M.-N. Garcia, M. Morel, *A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign*, Proc. LREC, pp. 2034-2037, Gênes, 2006.

[18] C. Benoît, *An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity*, Speech Commun., vol. 9, no. 4, pp. 293304, 1990.

[19] Noisex-92, Online, *http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.html*.

[20] J. Allen, D. Berkley, *Image method for efficiently simulating small-room acoustics*, JASA, vol. 65, no. 4, pp. 943-950, 1979.