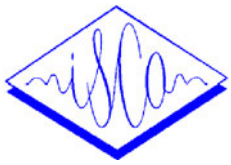
ISCA Archive
http://www.isca-speech.org/archive



7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Analysis and Synthesis of Hypo and Hyperarticulated Speech

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 analysis and synthesis of hypo and hyperarticulated speech in the framework of HMM-based speech synthesis. First of all, a new French database matching our needs was created, which contains three identical sets, pronounced with three different degrees of articulation: neutral, hypo and hyperarticulated speech. On that basis, acoustic and phonetic analyses were performed. It is shown that the degrees of articulation significantly influence, on one hand, both vocal tract and glottal characteristics, and on the other hand, speech rate, phone durations, phone variations and the presence of glottal stops. Finally, neutral, hypo and hyperarticulated speech are synthesized using HMM-based speech synthesis and both objective and subjective tests aiming at assessing the generated speech quality are performed. These tests show that synthesized hypoarticulated speech seems to be less naturally rendered than neutral and hyperarticulated speech.

Index Terms: Speech Synthesis, HTS, Speech Analysis, Expressive Speech, Voice Quality

1. Introduction

In this paper, we focus on the study of different speech styles, based on the degree of articulation: neutral speech, hypoarticulated (or casual) and hyperarticulated speech (or clear speech). 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 [1]. The influence of emotion on the articulation degree has been studied in [2], [3] and is out of the scope of this work.

The "H and H" theory [4] 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 [1]. Speakers can adopt a speaking style that allows them to be understood more easily in difficult communication situations. The degree of articulation is influenced by the phonetic context, the speech rate and 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 the differences between the real observations and the targets versus the speech rate. Because defining formant targets is not an easy task, Beller proposed in [1] a statistical measure of the degree of articulation by studying the joint evolution of the vocalic triangle area and the speech rate.

The goal of this study is to have a better understanding of the specific characteristics (acoustic and phonetic) governing hypo and hyperarticulated speech and to apply it to HMM synthesis. In order to achieve this goal, the paper is divided into two main parts: the analysis (Section 3) and synthesis (Section 4) of hypo and hyperarticulated speech.

In the first part, the acoustic (Section 3.1) and phonetic (Section 3.2) modifications are studied as a function of the degree of articulation. The acoustic analysis highlights evidence of both vocal tract and glottal characteristics changes, while the phonetic analysis focuses on showing evidence of glottal stops presence, phone variations, phone durations and speech rate changes. In the second part, the integration within a HMM-based speech synthesizer in order to generate the two degrees of articulation is discussed (Section 4.1). Both an objective and subjective evaluation are carried out with the aim of assessing how the synthetic speech quality is affected for both degrees of articulation. Finally Section 5 concludes the paper and some of our future works are given in Section 6.

2. Creation of a Database with various Degrees of Articulation

For the purpose of our research, a new French database was recorded 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, as neutral 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.

While recording hyperarticulated speech, the speaker was listening to a version of his voice modified by a "Cathedral" effect. This effect produces a lot of reverberations (as in a real cathedral), forcing the speaker to talk slower and as clearly as possible (more efforts to produce speech). On the other hand, while recording hypoarticulated speech, the speaker was listening to an amplified version of his own voice. This effect produces the impression of talking very close to someone in a narrow environment, allowing the speaker to talk faster and less clearly (less efforts to produce speech). Proceeding that way allows us to create a "standard recording protocol" to obtain repeatable conditions if required in the future.

3. Analysis of Hypo and Hyperarticulated Speech

3.1. Acoustic Analysis

Acoustic modifications in expressive speech have been extensively studied in the literature [7], [8], [9]. In the frame of this study, one can expect important changes related to the vocal tract function. Indeed, during the production of hypo and hyperarticulated speech, the articulatory strategy adopted by the speaker may dramatically vary. Although it is still not clear

whether these modifications consist of a reorganization of the articulatory movements, or of a reduction/amplification of the normal ones, speakers generally tend to consistently change their way of articulating. According to the "H and H" theory [4], speakers minimize their articulatory trajectories in hypoarticulated speech, resulting in a low intelligibility, while an opposite strategy is adopted in hyperarticulated speech. As a consequence, the vocal tract configurations may be strongly affected. The resulting changes are studied in Section 3.1.1.

In addition, the produced voice quality is also altered. Since voice quality variations are mainly considered to be controlled by the glottal source [9], Section 3.1.2 focuses on the modifications of glottal characteristics with regard to the degree of articulation.

3.1.1. Vocal Tract-based Modifications

In order to study the variations of the vocal tract resonances, the evolution of the vocalic triangle [1] with the degree of articulation was analyzed. This triangle consists of the three vowels $/a/$, $/i/$ and $/u/$ represented in the space of the two first formant frequencies $F1$ and $F2$ (here estimated via Wavesurfer [10]). For the three degrees of articulation, the vocalic triangle is displayed in Figure 1 for the original sentences. For information, ellipses of dispersion are also indicated on these plots. The first main conclusion is the significant reduction of the vocalic space as speech becomes less articulated. Indeed, as the articulatory trajectories are less marked, the resulting acoustic targets are less separated in the vocalic space. This may partially explain the lowest intelligibility in hypoarticulated speech. On the contrary, the enhanced acoustic contrast is the result of the efforts of the speaker under hyperarticulation. These changes of vocalic space are summarized in Table 1, which presents the area defined by the average vocalic triangles.

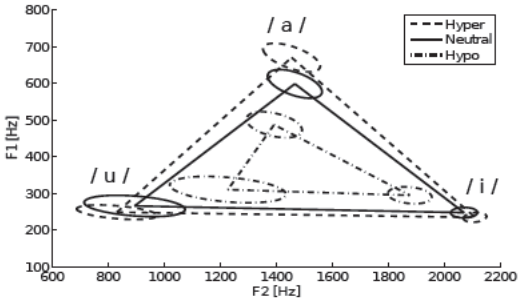


Figure 1: Vocalic triangle, for the three degrees of articulation, estimated on the **original** recordings. Dispersion ellipses are also indicated.

Dataset	Hyper	Neutral	Hypo
Original	0.274	0.201	0.059

Table 1: Vocalic space (in kHz^2) for the three degrees of articulation for the original sentences.

Inspecting the ellipses, it is observed that dispersion can be high for the vowel $/u/$, while data is relatively well concentrated for $/a/$ and $/i/$.

3.1.2. Glottal-based Modifications

As the most important perceptual glottal feature, pitch histograms are displayed in Figure 2. It is clearly noted that the more speech is articulated, the higher the fundamental frequency. Besides these prosodic modifications, we investigate how characteristics of the glottal flow are affected. In a first part, the glottal source is estimated by the Complex Cepstrum-based Decomposition algorithm (CCD, [12]). This method relies on the mixed-phase model of speech [13]. According to this model, speech is composed of both minimum-phase and maximum-phase components, where the latter contribution is only due to the glottal flow. By isolating the maximum-phase component of speech, the CCD method has shown its ability to efficiently estimate the glottal source. Using this technique, Figure 3 shows the averaged magnitude spectrum of the glottal source for the three degrees of articulation. First of all, a strong similarity of these spectra with models of the glottal source (such as the LF model [14]) can be noticed. Secondly it turns out that a high degree of articulation is reflected by a glottal flow containing a greater amount of high frequencies. Finally, it is also observed that the glottal formant frequency increases with the degree of articulation (see the zoom in the top right corner of Figure 3). In other words, the time response of the glottis open phase turns to be faster in hyperarticulated speech.

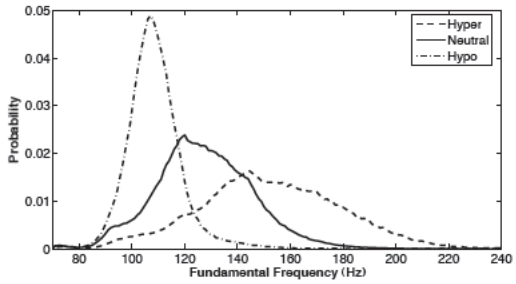


Figure 2: Pitch histograms for the three degrees of articulation.

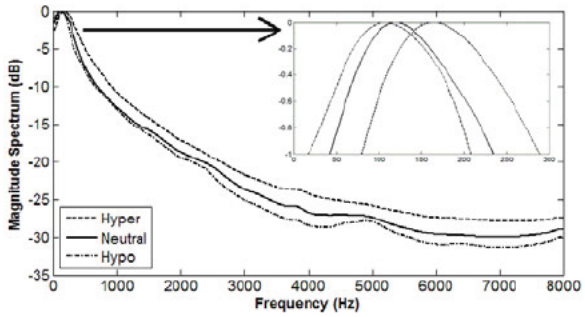


Figure 3: Averaged magnitude spectrum of the glottal source for the three degrees of articulation.

In a second part, the maximum voiced frequency is analyzed. In some approaches, such as the Harmonic plus Noise Model (HNM, [15]) or the Deterministic plus Stochastic Model of residual signal (DSM, [16]) which will be used for synthesis in Section 4, the speech signal is considered to be modeled by a non-periodic component beyond a given frequency. This maximum voiced frequency (F_m) demarcates the boundary between two distinct spectral bands, where respectively an harmonic and

a stochastic modeling (related to the turbulences of the glottal airflow) are supposed to hold. In this paper, F_m was estimated using the algorithm described in [15]. The corresponding histograms are illustrated in Figure 4 for the three degrees of articulation. It can be noticed from this figure that the more speech is articulated, the higher the F_m, the stronger the harmonicity, and consequently the weaker the presence of noise in speech. Note that the average values of F_m are respectively of 4215 Hz, 3950 Hz (confirming our choice of 4 kHz in [16]) and 3810 Hz for the three degrees of articulation.

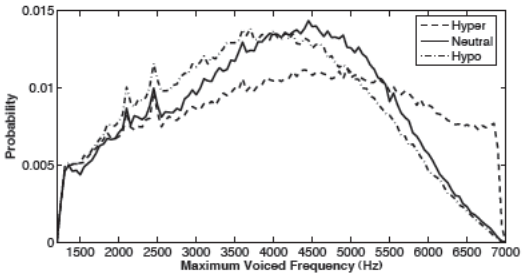


Figure 4: Histograms of the maximum voiced frequency for the three degrees of articulation.

3.2. Phonetic Analysis

Phonetic modifications in hypo and hyperarticulated speech are also very important characteristics to investigate. In the next paragraphs, glottal stops (Section 3.2.1), phone variations (Section 3.2.2), phone durations (Section 3.2.3) and speech rates (Section 3.2.4) are analyzed. In order to obtain reliable results, the entire database for each degree of articulation is used in this section. Moreover, the 36 standard French phones are considered ([25] from which /â/ and /ng/ are not used because they can be made from other phonemes, and /_/ is added). Note that results can vary from one speaker to another as pointed out in [1]. Eventually, the database was segmented using HMM forced alignment [26].

3.2.1. Glottal Stops

According to [17], a glottal stop is a cough-like explosive sound released just after the silence produced by the complete glottal closure. In French, such a phenomenon happens when the glottis closes completely before a vowel. A method for detecting glottal stops in continuous speech was proposed in [18]. However, this technique was not used here. Instead we detected glottal stops manually. Figure 5 shows, for each vowel, the number of glottal stops for each degree of articulation. It turns out from this figure that the number of glottal stops is much higher (almost always double) in hyperarticulated speech than in neutral and hypoarticulated speech (between which no sensible modification is noticed).

3.2.2. Phone Variations

Phone variations refer to phonetic insertions, deletions and substitutions that the speaker makes during hypo and hyperarticulation, compared to the neutral speech. This study has been performed at the phone level, considering the phone position in the word, and at the phone group level (groups of phones that were inserted, deleted or substituted).

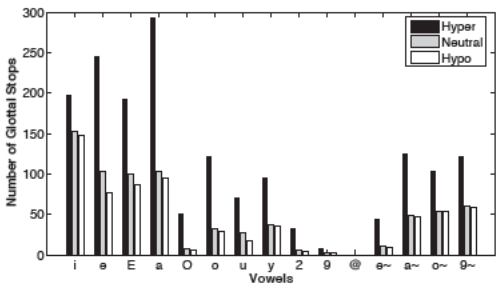


Figure 5: Number of glottal stops for each phone (vowel) and each degree of articulation.

For the sake of conciseness, only the most relevant differences will be given in this section. Table 2 shows, for each phone, the total proportion of phone deletions in hypoarticulated speech and phone insertions in hyperarticulated speech (first line). The position of these deleted/inserted phones inside the words are also shown: at the beginning (second line), in the middle (third line) and at the end (fourth line). Note that since there is no significant deletion process in hyperarticulation, no significant insertion process in hypoarticulation and no significant substitution process in both cases, they do not appear in Table 2.

In hyperarticulated speech, the only important insertions are breaks /_/ and Schwa /@/ (mostly at the end of the words). In hypoarticulated speech, breaks and Schwa (mostly at the end of the words) are often deleted, as /R/, /l/, /Z/ and /z/. Schwa, also called "mute e" or "unstable e", is very important in French. It is the only vowel that can or cannot be pronounced (all other vowels should be clearly pronounced), and several authors have focused on Schwa insertions and deletions in French. The analysis performed at the phone group level is still under development but we observed frequent phone group deletions in hypoarticulated speech (e.g. /R@/, /l@/ at the end of the words, */je suis/* (which means */I am/*) becoming */j'suis/* or even */chui/*, ...) and no significant group insertions in hyperarticulated speech. In both cases, no significant phone groups substitutions were observed.

3.2.3. Phone Durations

Intuitively, it is expected that the degree of articulation has an effect on phone durations, as well as on the speech rate (Section 3.2.4). Some studies are confirming that thought. In the approach exposed in [23], it is found evidence for the Probabilistic Reduction Hypothesis: word forms are reduced when they have a higher probability, and this should be interpreted as evidence that probabilistic relations between words are represented in the mind of the speaker. Similarly, [19] examines how that probability (lexical frequency and previous occurrence), speaking style, and prosody affect word duration, and how these factors interact.

In this work, we have investigated the phone duration variations between neutral, hypoarticulated and hyperarticulated speech. Vowels and consonants were grouped according to broad phonetic classes [25]. Figure 6 shows the histograms of (a) front, central, back and nasal vowels, (b) plosive and fricative consonants, and (c) breaks. Figure 7 shows the histograms of (a) semi-vowels and (b) trill consonants. As expected, one can see that, generally, phone durations are shorter in hypoar-

Phone		/j/	/H/	/t/	/k/	/z/	/Z/	/l/	/R/	/E/	/@/	/_/
Deletions (Hypoarticulation)	Tot	1.5	1.7	1.9	1.6	**3.1**	**5.1**	**2.2**	**3.4**	1.5	**29.7**	**14.2**
	Beg	0.14	0.57	0.14	0.38	0.0	**4.95**	0.26	0.03	0.89	11.49	**14.2**
	Mid	0.53	1.13	0.52	0.45	0.94	0.15	0.44	1.62	0.47	2.85	0.0
	End	0.82	0.0	1.24	0.77	**2.16**	0.0	**1.50**	**1.75**	0.14	**15.39**	0.0
Insertions (Hyperarticulation)	Tot	0.3	0.0	1.1	0.2	4.0	0.6	0.1	0.2	0.2	**40.0**	**26.5**
	Beg	0.0	0.0	0.10	0.0	0.0	0.0	0.025	0.0	0.05	0.60	**26.5**
	Mid	0.15	0.0	0.25	0.07	0.41	0.15	0.025	0.04	0.1	1.68	0.0
	End	0.15	0.0	0.75	0.13	3.59	0.45	0.05	0.16	0.05	**37.72**	0.0

Table 2: Total percentage (first line) of deleted and inserted phones in hypo and hyperarticulated speech respectively, and their repartition inside the words: beginning (second line), middle (third line), end (fourth line).

ticulation and longer in hyperarticulation. Breaks are shorter (and more rare) in hypoarticulation, but are as long as the ones in neutral speech and more present in hyperarticulation. An interesting characteristic of hypoarticulated speech is the concentration (high peaks) of semi-vowels and trill consonants in the short durations.

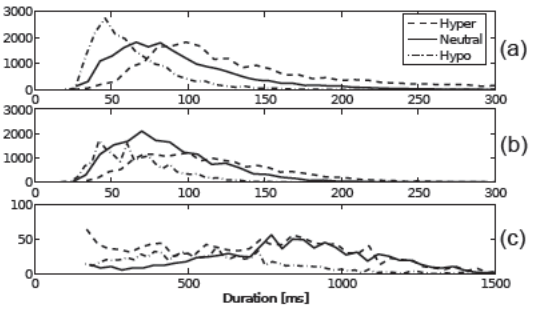


Figure 6: Phone durations histograms. (a) front, central, back & nasal vowels. (b) plosive & fricative consonants. (c) breaks.

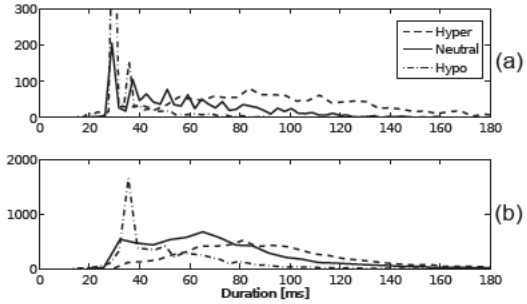


Figure 7: Phone durations histograms. (a) semi-vowels. (b) trill consonants.

3.2.4. Speech Rate

Speaking rate has been found to be related to many factors [20]. It is often defined as the average number of syllables uttered per second (pauses excluded) in a whole sentence [21], [22]. Based on that definition, Table 3 compares the three speaking styles.

As expected, hyperarticulated speech is characterized by a lower speech rate, a higher number of breaks (thus a longer pausing time), more syllables (final Schwa insertions), resulting in an increase of the total speech time.

Results	Hyper	Neutral	Hypo
Total speech time [s]	6076	4335	2926
Total syllable time [s]	5219	3618	2486
Total pausing time [s]	857	717	440
Total number of syllables	19736	18425	17373
Total number of breaks	1213	846	783
Speech rate [syllable/s]	3.8	5.1	7.0
Pausing time [%]	14.1	16.5	15.1

Table 3: Results for hypo, neutral & hyperarticulated speech.

On the other side, hypoarticulated speech is characterized by a higher speech rate, a lower number of breaks (thus a shorter pausing time), less syllables (final Schwa and other phone groups deletions), resulting in a decrease of the total speech time. An interesting property can be noted: because of the increase (decrease) in the total pausing time and the total speech time in hyper (hypo) articulated speech, the pausing time (thus the speaking time) expressed in percents of the total speech time is almost independent of the speech style.

4. Synthesis of Hypo and Hyperarticulated Speech

Synthesis of the articulation degree in concatenative speech synthesis has been performed in [5], by modifying the spectral shape of acoustic units according to a predictive model of the acoustic-prosodic variations related to the articulation degree. In this paper, we report our first attempts in synthesizing the two degrees of articulation of speech using HMM-based speech synthesis (via HTS [6]).

4.1. Integration within HMM-based Speech Synthesis

For each degree of articulation, a HMM-based speech synthesizer [24] was built, relying for the implementation on the HTS toolkit (version 2.1) publicly available in [6]. In each case, 1220 sentences sampled at 16kHz were used for the training, leaving around 10% of the database for synthesis. For the filter, we extracted the traditional Mel Generalized Cepstral coefficients (with frequency warping factor = 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 [16], since it was shown to significantly improve the naturalness of the delivered speech. More precisely, both deterministic and stochastic components of the DSM model were estimated from the training dataset for each degree of articulation. The spectral boundary between these two components was

chosen as the averaged value of the maximum voiced frequency described in Section 3.1.2.

The objective of this preliminary work was to assess the quality of the synthesized speech based only on phonetic transcription modifications. Therefore, hypo and hyperarticulated speech were obtained by manually modifying the phonetic transcriptions at the input of the synthesizer, according to Section 3.2.2 (our future natural language processor should do it automatically). In the following evaluations, original pitch and phone durations were imposed at the input of the synthesizers.

4.2. Acoustic Analysis

The same acoustic analysis as in Section 3.1.1 was performed on the sentences generated by the HMM-based synthesizer. Results are summarized in Figure 8 and in Table 4. Note the good agreement between vocalic spaces in original (see Section 3.1.1) and synthesized sentences.

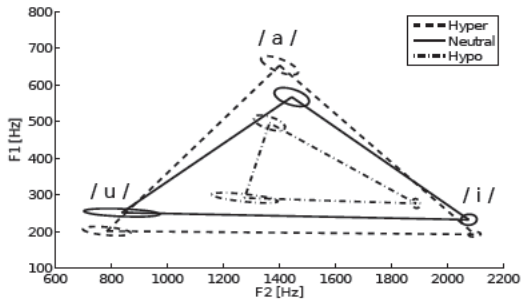


Figure 8: Vocalic triangle, for the three degrees of articulation, estimated on the sentences **generated** by the HMM-based synthesizer. Dispersion ellipses are also indicated.

Dataset	Hyper	Neutral	Hypo
Synthesis	0.299	0.201	0.063

Table 4: Vocalic space (in kHz^2) for the three degrees of articulation for the synthesized sentences.

The same conclusions as in Section 3.1.1 hold for the synthetic examples. In other words, the essential vocalic characteristics are preserved despite the HMM-based modeling and generation process. It can be however noticed that the dispersion of the formant frequencies is lower after generation, especially for $F1$. This is mainly due to an over-smoothing of the generated spectra (albeit the Global Variance method [11] was used).

4.3. Objective Evaluation

The goal of the objective evaluation is to assess whether HTS is capable of producing natural hypo and hyperarticulated speech and to which extent. The distance measure considered here is the mel-cepstral distortion between the target and the estimated mel-cepstra coefficients, expressed as:

$$Mel-CD = \frac{10}{ln(10)}\sqrt{2\sum_{d=1}^{24}(mc_d^{(t)} - mc_d^{(e)})^2} \qquad (1)$$

This mel-cepstral distortion is computed for all the vowels of the database. Table 5 shows the mean with its 95% confidence interval for each degree of articulation. This objective evaluation shows that the mel-cepstral distortion increases from hyper to hypoarticulated speech.

Results	Hyper	Neutral	Hypo
Mean $\pm$ CI	5.9 ± 0.1	6.3 ± 0.2	6.9 ± 0.1

Table 5: Objective evaluation results (in [dB]): mean score with its 95% confidence interval (CI) for each degree of articulation.

4.4. Subjective Evaluation

In order to confirm the objective evaluation conclusion, we performed a subjective evaluation. For this evaluation, the listener was asked to compare three sentences: A, the original; B, the sentence vocoded by DSM; C, the sentence synthesized by HTS using DSM as vocoder. He was asked to score, on a 9-point scale, the overall speech quality of B in comparison with A and C. B was allowed to vary from 0 (= same quality as A) to 9 (= same quality as C). Therefore this score should be interpreted in terms of a "distance" between B and A and C: the lower the score, the more B "sounds like" A and thus the better the quality, and conversely.

The test consists of 15 triplets (5 sentences per degree of articulation), giving a total of 45 sentences. Before starting the test, the listener was provided with some reference sentences covering most of the variations to help him familiarize with the scale. During the test, he was allowed to listen to the triplet of sentences as many times as he wanted, in the order he preferred (he was advised to listen to A and C before listening to B, in order to know the boundaries). However he was not allowed to come back to previous sentences after validating his decision.

The hypothesis made in this subjective evaluation is that the distance between A and B is constant, whatever the degree of articulation is. This hypothesis has been verified by informal listening. By proceeding this way, speech quality of C vs A can be assessed indirectly. 26 people, mainly naive listeners, participated to this evaluation. The mean score, corresponding to the "distance" between A and C, together with its 95% confidence interval for each articulation degree, on the 9-point scale, is shown in Figure 9. The lower the score, the more C "sounds like" A and thus the better the quality, and conversely. One can see that hypoarticulated speech is the worst, followed by neutral and hyperarticulated speech, therefore confirming the objective evaluation result.

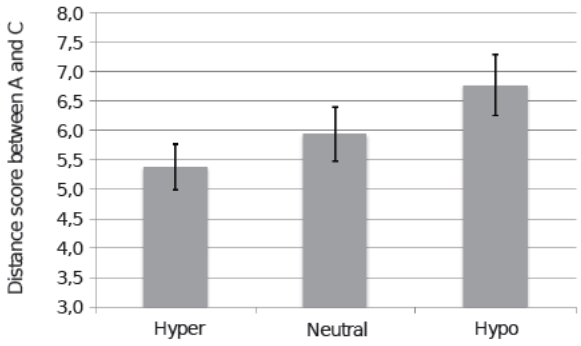


Figure 9: Subjective evaluation results: overall speech quality of the HMM-based speech synthesizer (mean score with its 95% confidence interval for each degree of articulation).

5. Conclusion

This work is a first approach towards HMM-based hyper and hypoarticulated speech synthesis. A new French database matching our needs was created: three identical sets, pronounced with three different degrees of articulation (neutral, hypo and hyperarticulated speech).

In a first step, acoustic and phonetic analyses were performed on these databases, and the influence of the articulation degree on various factors was studied. It was shown that hyperarticulated speech is characterized by a larger vocalic space (more efforts to produce speech, with maximum clarity), higher fundamental frequency, a glottal flow containing a greater amount of high frequencies and an increased glottal formant frequency, the presence of a higher number of glottal stops, breaks and syllables, significant phone variations (especially insertions), longer phone durations and lower speech rate. The opposite tendency was observed in hypoarticulated speech, except that the number of glottal stops was equivalent to the one in neutral speech and the significant phone variations were deletions.

In a second step, synthesizing hypo and hyperarticulated speech was performed using HTS, based on modifications of the phonetic transcriptions at the input of the synthesizer, and of the characteristics of the excitation modeling. Objective and subjective evaluations were proposed in order to assess the quality of the synthesized speech. These tests show that the worst speech quality was obtained for hypoarticulated speech.

Audio examples for each degree of articulation are available online via http://tcts.fpms.ac.be/~picart/HypoAndHyperarticulatedSpeech_Demo.html.

6. Discussion and Future Works

The ultimate goal of our research is to be able to synthesize hypo and hyperarticulation, directly from an existing neutral voice (using voice conversion), without requiring recordings of new hypo and hyperarticulated databases (as done in this work). Right now, as the objective and subjective evaluations showed, the HMM-based speech synthesizers are not able to synthesize hypo and hyperarticulated speech with the same quality, even using the real hypo and hyperarticulated databases. It is therefore worth focusing on improving the current synthesis method before starting the next step: speaking style conversion. We will first investigate the simple methods for improving speaker-similarity in HMM-based speech synthesis proposed by [27].

7. 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). Authors would like to thank Y. Stylianou for providing the algorithm for extracting F_m, and also Acapela Group SA for providing useful advices on database recordings and helping us in segmenting the database.

8. References

[1] 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.

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

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

[5] J. Wouters, *Analysis and Synthesis of Degree of Articulation*, PhD Thesis, Katholieke Universiteit Leuven (KUL), Belgium, 1996.

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

[7] D. Klatt, L. Klatt, *Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers*, JASA, vol. 87, pp. 820-857, 1990.

[8] D. Childers, C. Lee, *Vocal Quality Factors: Analysis, Synthesis, and Perception*, JASA, vol. 90, pp. 2394-2410, 1991.

[9] E. Keller, *The analysis of voice quality in speech processing*, Lecture Notes in Computer Science, pp. 54-73, 2005.

[10] K. Sjolander, J. Beskow, *Wavesurfer - an open source speech tool*, ICSLP, vol.4, pp. 464-467, 2000.

[11] T. Toda, K. Tokuda, *A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis*, IEICE Trans. on Information and Systems, vol. E90-D, pp. 816-824, 2007.

[12] T. Drugman, B. Bozkurt, T. Dutoit, *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*, Interspeech Conference, 2009.

[13] B. Bozkurt, T. Dutoit, *Mixed-phase speech modeling and formant estimation, using differential phase spectrums*, VOQUAL’03, pp. 21-24, 2003.

[14] G. Fant, J. Liljencrants, Q. Lin, *A four parameter model of glottal flow*, STL-QPSR4, pp. 1-13, 1985.

[15] Y. Stylianou, *Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis*, IEEE Trans. Speech and Audio Processing, vol. 9(1), pp. 21-29, 2001.

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

[17] [Online] Glottal Stop: http://www.bbc.co.uk/dna/h2g2/A1002808

[18] B. Yegnanarayana, S. Rajendran, Hussien Seid Worku, Dhananjaya N., *Analysis of Glottal Stops in Speech Signals*, Proc. Interspeech, 2009.

[19] R. E. Baker, A. R. Bradlow, *Variability in Word Duration as a Function of Probability, Speech Style, and Prosody*, Language and Speech, Vol. 52, No. 4, 391-413, 2009.

[20] J. Yuan, M. Liberman, C. Cieri, *Towards an Integrated Understanding of Speaking Rate in Conversation*, Interspeech 2006, 541-544, Pittsburgh, PA, 2006.

[21] G. Beller, T. Hueber, D. Schwarz, X. Rodet, *Speech Rates in French Expressive Speech*, Third International Conference on Speech Prosody, Dresden, Germany, 2006.

[22] S. Roekhaut, J-P. Goldman, A. C. Simon, *A Model for Varying Speaking Style in TTS systems*, Fifth International Conference on Speech Prosody, Chicago, IL, 2010.

[23] D. Jurafsky, A. Bell, M. Gregory, W. D. Raymond, *Probabilistic Relations between Words: Evidence from Reduction in Lexical Production*, in Bybee, Joan and Paul Hopper (eds.). Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins. 229-254. 2001.

[24] H. Zen, K. Tokuda, A. W. Black, *Statistical parametric speech synthesis*, Speech Communication, Volume 51, Issue 11, November 2009, Pages 1039-1064, 2009.

[25] [Online] Phonetic: http://phonetique.free.fr/api.pdf

[26] F. Malfrere, O. Deroo, T. Dutoit, C. Ris, *Phonetic alignement : speech-synthesis-based versus Viterbi-based*, Speech Communication, vol. 40, n4, pp. 503-517, 2003.

[27] J. Yamagishi, S. King, *Simple Methods for Improving Speaker-Similarity of HMM-based Speech Synthesis*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, Texas, 2010.