Analysis and HMM-based Synthesis of Hypo and Hyperarticulated Speech

Benjamin Picart*, Thomas Drugman, Thierry Dutoit

TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Belgium

Abstract

Hypo and hyperarticulation refer to the production of speech with respectively a reduction and an increase of the articulatory efforts compared to the neutral style. Produced consciously or not, these variations of articulatory efforts depend upon the surrounding environment, the communication context and the motivation of the speaker with regard to the listener. The goal of this work is to integrate hypo and hyperarticulation into speech synthesizers, such that they are more realistic by automatically adapting their way of speaking to the contextual situation, like humans do. Based on our preliminary work, this paper provides a thorough and detailed study on the analysis and synthesis of hypo and hyperarticulated speech. It is divided into three parts. In the first one, we focus on both acoustic and phonetic modifications due to articulatory effort changes. The second part aims at developing a HMM-based speech synthesizer allowing a continuous control of the degree of articulation. This requires to first tackle the issue of speaking style adaptation to derive hypo and hyperarticulated speech from the neutral synthesizer. Once this is done, an interpolation and extrapolation of the resulting models enables to finely tune the voice so that it is generated with the desired articulatory efforts. Finally the third and last part focuses on a perceptual study of speech with a variable articulation degree, where it is analyzed how intelligibility and various other voice dimensions are affected.

Keywords: Speech Synthesis, HTS, Speech Analysis, Expressive Speech,

[*]Corresponding author. Tel. +32 65 374746. Fax +32 65 374729

Email addresses: `benjamin.picart@umons.ac.be` (Benjamin Picart), `thomas.drugman@umons.ac.be` (Thomas Drugman), `thierry.dutoit@umons.ac.be` (Thierry Dutoit)

1. Introduction

The "H and H" theory [Lindblom (1983)] proposes two degrees of articulation of speech: hyperarticulated speech, for which speech clarity tends to be maximized by increasing the articulation efforts to produce speech, and hypoarticulated speech, where the speech signal is produced with minimal articulation efforts. Therefore the degree of articulation (DoA) provides information on the motivation and personality of the speaker vs. the listeners [Beller (2009)]. Indeed, when talkers speak, they also listen [Cooke et al. (2012)]. Speakers can adopt a speaking style allowing them to be understood more easily in difficult communication situations. In this work, "hyperarticulated speech" (HPR) refers to the situation of a teacher or a speaker talking in front of a large audience (important articulation efforts have to be made to be understood by everybody). "Hypoarticulated speech" (HPO) 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). "Neutral speech" (NEU) refers to the daily life situation of a person reading aloud a text emotionless (e.g. no happiness, no anger, no excitement, etc) and without any specific articulation efforts to produce the speech, keeping only the sentence intonation: pitch rise for a question, flat pitch for an affirmative or negative sentence, etc. It is worth noting that these three modes of expressivity are emotionless, but can vary amongst speakers as reported in [Beller (2009)]. The influence of emotion on the DoA has been studied in [Beller (2007)] [Beller et al. (2008)] and is out of the scope of this work.

The DoA 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 DoA 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 [Beller (2009)] a statistical measure of the DoA by studying the joint evolution of the vocalic triangle (i.e. the shape formed by the vowels $/a/$, $/i/$ and $/u/$ in the $F1$ - $F2$ space) area and the speech rate. A recent study presented a computational model of human speech production to provide a continuous adjustment according to environmental conditions [Nicolao et al. (2012)].

In direct connection with HPR speech, the “Lombard effect” [Lombard (1911)] refers to the speech changes due to the immersion of the speaker in a noisy environment. It is indeed known that a speaker tends to increase his vocal efforts to be more easily understood while talking in a background noise [Summers et al. (1988)]. Various aspects of the Lombard effect were already studied, including acoustic and articulatory characteristics [Garnier et al. (2006a)] [Garnier et al. (2006b)], features extracted from the glottal flow [Drugman and Dutoit (2010)], or changes of F0 and of the spectral tilt [Lu and Cooke (2009)], etc.

Some works have been done in the framework of concatenative speech synthesis to enhance the speech intelligibility by means of Lombard or HPR speech. For example, speech intelligibility improvement has been carried out for a limited domain task in [Langner and Black (2005)] based on voice conversion techniques. For this, they recorded the CMU_SIN database [Langner and Black (2004)] containing two parallel corpora obtained respectively under clean and noisy conditions. Another example is the Loudmouth synthesizer [Patel et al. (2006)], which emulates human modifications (both acoustic and linguistic) to speech in noise by manipulating word duration, fundamental frequency and intensity. In [Bonardo and Zovato (2007)], it is proposed to tune dynamic range controllers (e.g. compressors and limiters) and some user controls (e.g. speaking rate and loudness) to improve the intelligibility of synthesized speech. Various methods allowing automatic modification of speech in order to achieve the same goal are investigated in [Anumanchipalli et al. (2010)] (e.g. boosting the signal amplitude in important frequency bands, modification of prosodic and spectral properties, etc). Another work [Cerňak (2006)] introduced an additional measure evaluating intelligibility for the unit cost in unit selection based speech synthesis.

A new method for extracting or modifying mel cepstral coefficients based on an intelligibility measure for speech in noise, the Glimpse proportion measure, has been proposed in [Valentini-Botinhao et al. (2012a)] [Valentini-Botinhao et al. (2012b)]. Lombard speech synthesis in HMM-based speech synthesis has also been performed in [Raitio et al. (2011)]. Nonetheless, contrarily to the Lombard effect which is a reflex produced unconsciously due to the noisy surrounding environment [Junqua (1993)], HPR speech is defined as the voice produced with increased articulatory efforts compared to the NEU style. From a general point of view, these latter efforts might therefore also result from a voluntary decision to enhance speech intelligibility to facilitate the listener’s comprehension (like in the case of teaching). A similar case happens when people hyperarticulate in front of interactive systems, hoping to correct their recognition errors [Oviatt et al. (1998)].

This article provide a detailed and complete study on the integration of the DoA in HMM-based speech synthesis, based on our preliminary works on the subject [Picart et al. (2010)] [Picart et al. (2011a)] [Picart et al. (2011b)] [Picart et al. (2012)]: NEU speech, HPO (or casual) and HPR (or clear) speech. HPO and HPR speech are of interest in many daily life 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 HPO speech), improving intelligibility performance in adverse environments (e.g. GPS voice inside a moving car, train or flight information in stations or halls), adapting the difficulty level when learning foreign languages with the student's progresses (i.e. from HPR to HPO speech), etc. Note also that the ultimate goal of our research is to be able to continuously control the DoA of an existing standard neutral voice for which no HPO and HPR recordings are available. The results of this article are therefore necessary and essential to reach this objective.

For this, the article is divided into three main parts. Based on a database recorded specifically for this work (Section 2) and which contains three parallel corpora (one for each DoA to be studied - NEU, HPO and HPR speech), the first part focuses on the analysis of the effects induced by the DoA on the speech signal (Section 3). This is performed both at the acoustic (Section 3.1) and phonetic (Section 3.2) levels, in order to have a better understanding of the specific characteristics governing HPO and HPR speech.

The second part is devoted to the integration of the DoA in the HMM-based speech synthesis framework (Section 4), which can be subdivided into three tasks: *i)* training of a HMM-based speech synthesizer, using the whole database described in Section 2, for each DoA considered in this work (NEU, HPO and HPR) in Section 4.1; *ii)* being able to produce HPO and HPR speech directly from the NEU synthesizer, by studying the efficiency of speaking style adaptation as a function of the size of the adaptation database (Section 4.2); *iii)* implementing a continuous control (also called tuner) of the DoA, manually adjustable by the user, to obtain not only NEU, HPO and HPR speech, but also any intermediate, interpolated or extrapolated DoA in a continuous way (Section 4.3). Speaker adaptation [Yamagishi et al. (2009)] 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 [Tachibana et al. (2003)] [Nose et al. (2009)].

The third part targets a perceptual multi-dimensional assessment of the DoA of the synthesizers (Section 5). We first evaluate the necessity of integrating a variable DoA in a HMM-based speech synthesis system when this latter is embedded in adverse conditions, which happens very often in daily life applications: for example, GPS voice inside a moving car (additive noise), train or flight information in stations or halls (reverberation), etc. The intelligibility of generated voice is studied as a function of the DoA, as well as the type and level of degradation. Secondly, the effectiveness of synthesized speech with variable articulatory efforts is compared to the original recordings through 7 aspects: overall quality, comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation. Finally, Section 6 concludes this research work by summarizing its major accomplishments.

2. Database with various Degrees of Articulation

For the purpose of our research, a new French database was recorded. It consists of utterances produced by a single male speaker, aged 25 and native French (Belgium) speaking. The database contains three separate sets, each set corresponding to one DoA (NEU, HPO and HPR). For each set, the speaker was asked to pronounce the same 1359 phonetically balanced sentences (around 75, 50 and 100 minutes of NEU, HPO and HPR speech respectively), as emotionless as possible. The speaker was placed inside a sound-proof room, equipped with a screen displaying the sentences to be pronounced, and with an AKG C3000B microphone. The audio acquisition system Motu 8pre was used outside this room, with a sampling rate of 44.1 kHz. Finally, a headset was provided to the speaker for both HPO and HPR recordings, in order to induce him to speak naturally while modifying his DoA.

While recording HPR 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 articulatory efforts to produce speech). The “Cathedral” environment was generated by the digital multi-effects processor Behringer Virtualizer DSP1000. On the other hand, while recording HPO 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 articulatory efforts to produce speech). The amplification effect was created using the Powerplay Pro-8 HA8000 amplifier. Proceeding that way allows us to create a "standard recording protocol" to obtain repeatable conditions if required in the future.

Recordings were then resampled to 16 kHz and normalized in loudness. This is because we want our study to be independent of the level of energy of speech, and focus on the phonetic and prosodic modifications, as well as in the acoustic changes related to the vocal tract function and to the glottal production. Finally, utterances have been segmented such that they start and finish with a silence of about 200 ms.

3. Acoustic and Phonetic Features in Hypo and Hyperarticulated Speech

One can expect important changes during the production of HPO and HPR speech, compared to NEU speech style. For example, [Oviatt et al. (1998)] provide, amongst others, a comprehensive analysis of acoustic, prosodic, and phonological adaptations to speech during human-computer error resolution. These modifications could be categorized in two main parts: acoustic (Section 3.1) and phonetic (Section 3.2) variations. The first part is related to the speech production using the vocal tract (Section 3.1.1) and the glottal excitation (Section 3.1.2), while the second part focuses on the changes induced on the phonetic transcriptions. The latter section analyses respectively glottal stops (Section 3.2.1), phone variations (Section 3.2.2), phone durations (Section 3.2.3), and the speech rate (Section 3.2.4) for each DoA. Note that all the results we report throughout this section were obtained by an analysis led on the entire original corpora (as described in Section 2).

3.1. Acoustic Analysis

Acoustic modifications in expressive speech have been extensively studied in the literature [Klatt and Klatt (1990)], [Childers and Lee (1991)], [Keller (2005)]. Important changes related to the response of the vocal tract (also referred to as supralaryngeal structures in articulatory phonetics [Laver (1994)]) are expected in this study. Indeed, the articulatory strategy adopted by the speaker may dramatically vary during the production of HPO and HPR speech. Although it is still not clear whether these modifications consist of a reorganization of the articulatory movements, or of a reduction or an amplification of the normal ones, speakers generally tend to consistently change their way of articulating. According to the “H and H” theory [Lindblom (1983)], speakers minimize their articulatory trajectories in HPO speech, resulting in a low intelligibility, while an opposite strategy is adopted in HPR speech. As a consequence, vocal tract (or supralaryngeal) 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 [Drugman et al. (2012)] [D’Alessandro (2006)] [Södersten et al. (1995)], Section 3.1.2 focuses on the modifications of glottal characteristics (also sometimes called laryngeal features [Laver (1994)]) with regard to the DoA.

3.1.1. Vocal Tract-based Modifications

Beller analyzed in [Beller (2009)] the evolution of the vocalic triangle with the DoA, providing interesting information about the variations of the vocal tract resonances. The vocalic triangle is the shape formed by the vowels $/a/$, $/i/$ and $/u/$ in the space constituted by the two first formant frequencies $F1$ and $F2$ (here estimated via the Wavesurfer software [Sjolander and Beskow (2000)]). Figure 1 displays, for the original sentences, the evolution of the vocalic triangle with the DoA. Dispersion ellipses are also indicated on this figure for information. It is observed that dispersion can be high for the vowel $/u/$ (particularly for F2), while data are relatively well concentrated for vowels $/a/$ and $/i/$. A significant reduction of the vocalic triangle area is clearly noticed as speech becomes less articulated: from HPR ($0.274\ kHz^2$) to NEU ($0.201\ kHz^2$) to HPO speech ($0.059\ kHz^2$). As a consequence of this reduction, acoustic targets become less separated in the vocalic space, confirming that articulatory trajectories are less marked during an HPO strategy. This explains also partially the lowest intelligibility in HPO speech. The opposite tendency is observed for HPR speech, resulting from the increased articulatory efforts produced by the speaker.

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 on average and the more important its dynamics range. Besides these prosodic modifications, we investigated how

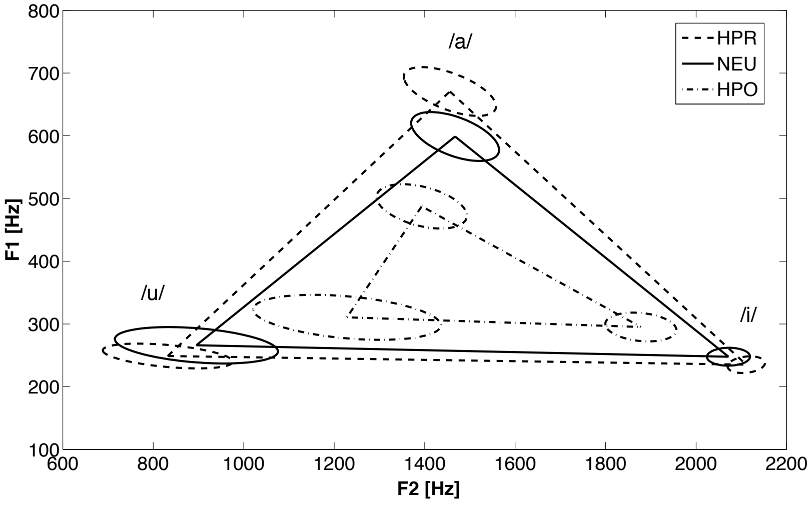

Figure 1: Vocalic triangle estimated on the original recordings for each DoA, together with dispersion ellipses.

some characteristics of the glottal flow are affected. The interested reader is referred to [Picart et al. (2010)] for more details about the following results.

The glottal source is estimated by the Complex Cepstrum-based Decomposition algorithm (CCD, [Drugman et al. (2011)]) as it was shown in [Drugman et al. (2012)] to provide the best results of the glottal source estimation. This technique relies on the mixed-phase model of speech [Bozkurt and Dutoit (2003)]. 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. The averaged magnitude spectrum of the glottal source for each DoA was computed using this technique, on the original data contained in our database (Section 2), and our conclusive observations are the following:

- resulting spectra have a strong similarity with models of the glottal source (such as the LF model [Fant et al. (1985)]), which corroborates the validity of these models and of the estimation process;
- a high DoA is characterized by a glottal flow containing more energy in the high frequencies, compared to the NEU case;
- the glottal formant frequency increases with the DoA, meaning that

the glottal open phase is more abrupt in HPR speech.

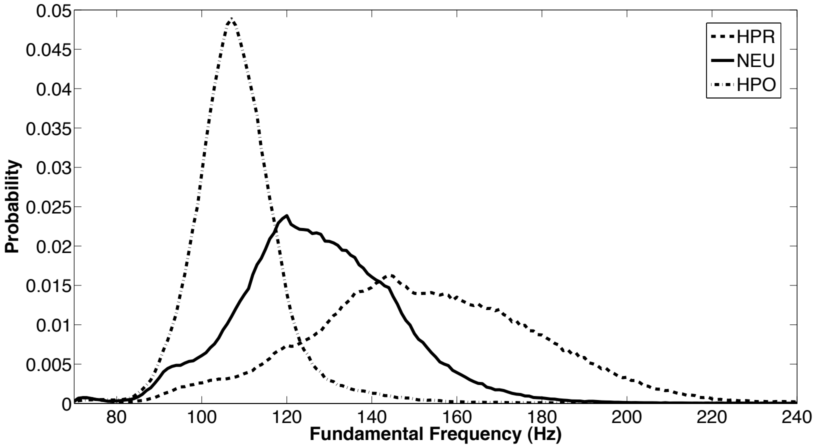

Figure 2: Pitch histograms for each DoA.

In some approaches, such as the Harmonic plus Noise Model (HNM, [Stylianou (2001)]) or the Deterministic plus Stochastic Model of residual signal (DSM, [Drugman and Dutoit (2012)]) 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. From our experiments, it turns out that:

- the more speech is articulated, the higher F_m, the stronger the harmonicity, and consequently the weaker the presence of noise in speech;
- the average values of F_m are 4215 Hz (HPR), 3950 Hz (NEU) and 3810 Hz (HPO). Note that this confirms the choice of 4 kHz for the synthesis of NEU speech in [Pantazis and Stylianou (2008)] or [Drugman and Dutoit (2012)].

3.2. Phonetic Analysis

In complement to the acoustic analysis of HPO and HPR speech in Section 3.1, we also investigate their phonetic modifications compared to NEU style. In the following, 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 studied. These results are here reported for the whole database described in Section 2, although such phonetic changes are known to have a certain inter-speaker variability [Beller (2009)]. Note that the database was segmented using HMM forced alignment [Malfrere et al. (2003)] using the 36 standard French phones and the SAMPA phonetic alphabet.

3.2.1. Glottal Stops

A glottal stop [Gordon and Ladefoged (2001)] [Borroff (2007)] 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 [Yegnanarayana et al. (2008)]. However, we chose to detect glottal stops manually in this study. The amount of glottal stops for each vowel is displayed in Figure 3, for each DoA. It interestingly appears that HPR speech is characterized by a higher amount of glottal stops (almost always double) than NEU and HPO speech, between which no sensible modification is noticed. This is an expected characteristic of the DoA. Indeed, HPR speech aims at increasing the intelligibility of a message, compared to NEU style, requiring more articulatory efforts. For example, word emphasis highlighting important information can be produced by the speaker under HPR by inserting a short break before the emphasis, producing at the same time a glottal stop.

3.2.2. Phone Variations

Compared to NEU speech style, any phonetic insertion, deletion and substitution made by the speaker under HPO and HPR are part of the phone variations. This study has been performed both at the phone level, considering the phone position in the word, and at the phone group level, considering groups of phones that were inserted, deleted or substituted.

Table 1 presents the total proportion of phone deletions in HPO speech and phone insertions in HPR speech (first row) for each phone (the most significant results are highlighted). The position of these deleted and inserted phones inside the words are also indicated: at the beginning (second row), in the middle (third row) and at the end (fourth row). Note that only the phones with the most relevant differences are shown in this table for the sake of conciseness (see [Picart et al. (2010)] for more results). Note also that

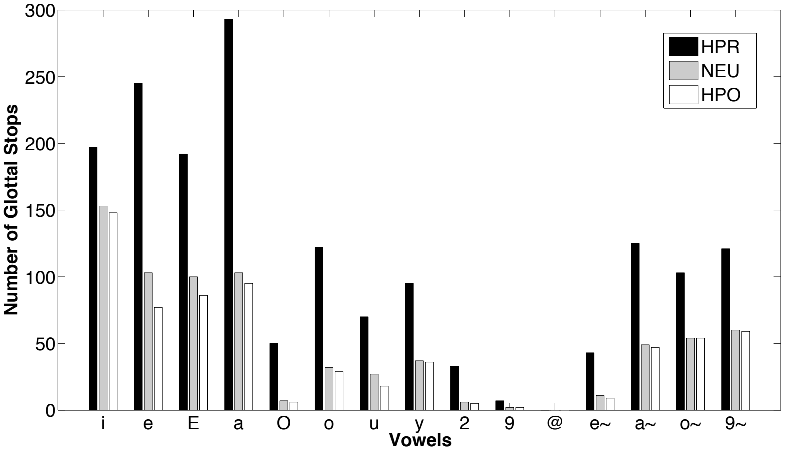

Figure 3: Number of glottal stops for each vowel and each DoA.

since there is no significant deletion process in HPR, no significant insertion process in HPO and no significant substitution process in both cases, they are not shown in Table 1.

Phone		/z/	/Z/	/l/	/R/	/@/	/_/
Deletions (HPO)	Tot	3.1	**5.1**	2.2	3.4	**29.7**	**14.2**
	Beg	0.0	4.95	0.26	0.03	11.49	14.2
	Mid	0.94	0.15	0.44	1.62	2.85	0.0
	End	2.16	0.0	1.50	1.75	15.39	0.0
Insertions (HPR)	Tot	4.0	0.6	0.1	0.2	**40.0**	**26.5**
	Beg	0.0	0.0	0.025	0.0	0.60	26.5
	Mid	0.41	0.15	0.025	0.04	1.68	0.0
	End	3.59	0.45	0.05	0.16	37.72	0.0

Table 1: Deleted and inserted phones percentage in HPO and HPR speech respectively, compared to NEU style, and their repartition inside the words: total (first row), beginning (second row), middle (third row), end (fourth row).

The most important variations concern breaks /_/ and Schwa /@/ deletions for HPO speech and insertions for HPR speech. Moreover, HPO speech counts also other significant phone deletions, i.e. /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 its use in French (see for example [Browman and Goldstein (1994)], [Adda-Decker et al. (1999)]). Besides, it is widely used by French speakers to mark hesitations. These conclusions with the phone Schwa are therefore probably specific to French and an extension of this phenomenon to other languages would therefore require further study.

The analysis performed at the phone group level revealed similar tendencies. While no significant group insertions in HPR speech have been detected, frequent phone group deletions in HPO speech were found: e.g. $/R@/$, $/l@/$ at the end of the words, "je suis" (which means "I am") becoming "j'suis" or even "chui", etc. However, no significant phone groups substitutions were observed in either cases.

3.2.3. Phone Durations

It is intuitively expected that the DoA influences phone durations, since HPO and HPR speech respectively target different intelligibility goals. This will directly affect the speech rate (Section 3.2.4). Some studies confirm that thought. Evidences for the *Probabilistic Reduction Hypothesis* are explained in the approach exposed in [Jurafsky et al. (2001)]: 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, [Baker and Bradlow (2009)] examines how that probability (lexical frequency and previous occurrence), speaking style, and prosody affect word duration, and how these factors interact between each others.

Phone duration variations between NEU, HPO and HPR speech were studied in [Picart et al. (2010)]. Vowels and consonants were grouped according to broad phonetic classes. We showed in that study that durations of front, central, back and nasal vowels are shorter during HPO and longer under HPR speech on average. The same conclusion was also true for plosive and fricative consonants. HPO speech contains shorter and fewer breaks, while HPR speech involves more of them, as long as those in NEU speech. We also interestingly noted a very high number of short-duration (around 30 ms) semi-vowels and trill consonants in HPO speech.

3.2.4. Speech Rate

Speaking rate has been found to be related to many factors [Yuan et al. (2006)]. It is often defined as the average number of syllables uttered per second (pauses excluded) in a whole sentence [Beller et al. (2006)] [Roekhaut et al. (2010)]. Table 2 compares the speaking styles corresponding to each DoA, following the previous definition.

Results	HPR	NEU	HPO
Total speech time [s]	6076 (+ 40.2 %)	4335	2926 (- 32.5 %)
Total syllable time [s]	5219 (+ 44.3 %)	3618	2486 (- 31.3 %)
Total pausing time [s]	857 (+ 19.5 %)	717	440 (- 38.6 %)
Total number of syllables	19736 (+ 7.1 %)	18425	17373 (- 5.7 %)
Total number of breaks	1213 (+ 43.4 %)	846	783 (- 7.4 %)
Speech rate [syllable/s]	3.8 (- 25.5 %)	5.1	7.0 (+ 37.3 %)
Relative pausing time [%]	14.1 (- 14.5 %)	16.5	15.1 (- 8.5 %)

Table 2: Speech rates and related time information for NEU, HPO & HPR speech, together with the positive or negative variation from the NEU style (in [%]).

As expected, HPR speech is characterized by a lower speech rate, a higher number of breaks (thus a longer pausing time), more syllables (due to final Schwa insertions in particular), resulting in an increase of the total speech time. On the other side, HPO speech is characterized by a higher speech rate, a lower number of breaks (thus a shorter pausing time), less syllables (due to final Schwa and other phone groups deletions), resulting in a decrease of the total speech time.

An interesting property can be noted: since both the total pausing time and the total speech time vary in about the same proportion (increase in HPR speech and decrease in HPO speech), the relative pausing time (and consequently the relative speaking time) percentage is almost independent of the speaking style. It seems that a speaker modifying his DoA controls unconsciously the relative proportions of speech and pausing time.

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

Only a few studies have been conducted on the synthesis of the DoA. Wouters made the first attempts within the context of concatenative speech synthesis [Wouters (2001)], by modifying the spectral shape of acoustic units according to a predictive model of the acoustic-prosodic variations related to the DoA. In this work, we focus on the synthesis of the DoA in the context of statistical parametric speech synthesis, using the HMM-based Speech Synthesis System HTS [Zen et al. (2009)].

First of all, a specific HMM-based speech synthesizer is built for each DoA (NEU, HPO and HPR), using the databases described in Section 2. The efficiency of speaking style adaptation to produce NEU, HPO and HPR speech directly from the NEU synthesizer is then studied. Finally, the implementation of a continuous control (tuner) of the DoA on this NEU synthesizer is detailed, with the goal of obtaining any interpolated or extrapolated DoA in a continuous way.

4.1. Speaker-Independent Full Data Model Training

4.1.1. Method

Relying on the implementation of the HTS toolkit[1] (version 2.1) publicly available, a HMM-based speech synthesizer [Zen et al. (2009)] was built for each DoA (NEU, HPO and HPR). Each database recorded as explained in Section 2 was used for training the corresponding synthesizer. A common practice when dealing with a database consists in keeping 90% of the data for training the models and leaving the rest for testing (here for synthesis). Therefore, for each DoA, 1220 sentences sampled at 16 kHz were used for the training (called the *training set*), leaving around 10% of the database for synthesis (called the *synthesis set*).

The speech signal is modeled and vocoded by a source-filter approach. The filter is represented by the Mel Generalized Cepstral (MGC [Tokuda et al. (1994)]) coefficients (with $\alpha = 0.42$, $\gamma = 0$ and order of MGC analysis = 24), and the excitation signal is based upon the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in [Drugman and Dutoit (2012)]. This model was shown to significantly increase the naturalness of the produced speech. More precisely, both deterministic and stochastic components of DSM were estimated on the training dataset for each DoA. The spectral boundary between these two components was fixed as the averaged value of the maximum voiced frequency described in Section 3.1.2. Note also that our version of HTS used 75-dimensional MGC parameters (including Δ and Δ^2), and each covariance matrix of the state output and state duration distributions were diagonal.

[1] http://hts.sp.nitech.ac.jp/

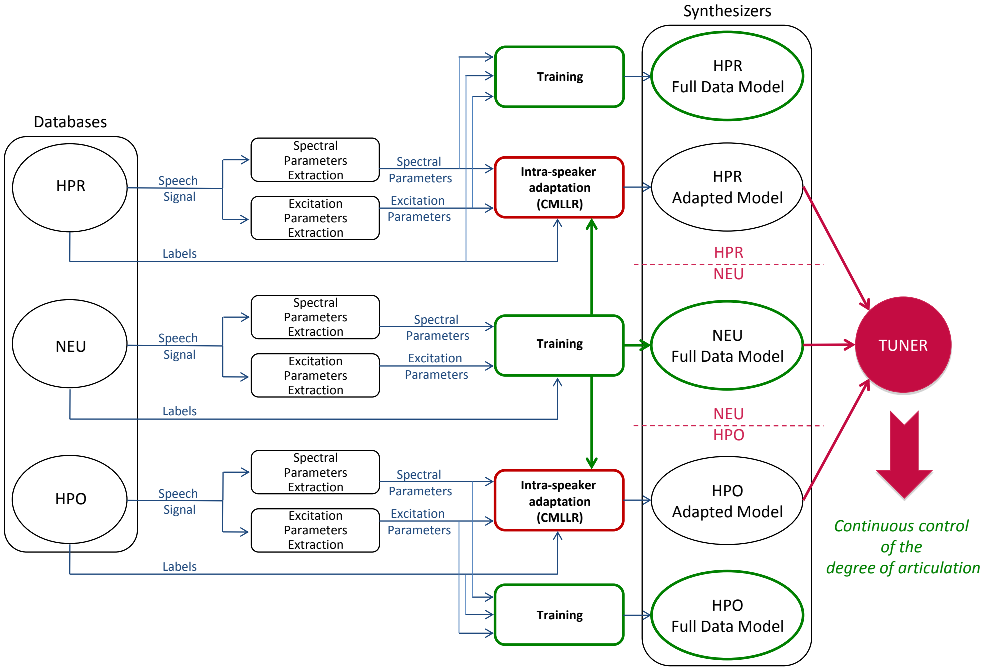

Figure 4: Standard training of the NEU, HPO and HPR full data models (Section 4.1.1), from the database containing 1220 training sentences for each DoA. Adaptation of the NEU full data model using CMLLR transform with HPO and HPR speech data to produce HPO and HPR adapted models (Section 4.2.1). Implementation of a tuner, manually adjustable by the user, for a continuous control of the DoA (Section 4.3.1).

Since the three synthesizers implemented at this point are trained on the entire training sets, they will be referred to as *full data models* in the following of this work. Figure 4 shows the general architecture of our system. At this point of the text, only the full data models training should be considered. The following evaluations are led on the synthesis set of the database.

4.1.2. Objective Evaluation

An objective evaluation is first conducted in order to assess the quality of the full data models. Yamagishi proposed in [Yamagishi and Kobayashi (2007)] for this the use of the following three objective measures: the average Mel-Cepstral Distortion (MCD, expressed in decibel), the Root-Mean-Square Error (RMSE) of log F0 (RMSE_lf0, expressed in cent), the RMSE of vowel durations (RMSE_dur, expressed in terms of number of frames). These measures reflect differences regarding three complementary aspects of speech.

The RMSE_lf0 is obviously computed for regions where both the original recordings and the full data models are voiced, since log F0 is not observed in unvoiced regions. Cent is a logarithmic unit used for musical intervals (100 cents correspond to a semitone, twelve semitones correspond to an octave, which means doubling of the frequency).

The MCD between the target and the estimated mel-cepstra coefficients (noted respectively $mc_d^{(t)}$ and $mc_d^{(e)}$, and computed from the original and synthesized versions of the same utterance) is expressed as:

$$MCD = \frac{10}{ln(10)}\sqrt{2\sum_{d=1}^{25}(mc_d^{(t)} - mc_d^{(e)})^2} \quad (1)$$

Target and estimated frames should have a one-to-one correspondence in order to compute an objective distance, which could either be for cepstrum or pitch with dedicated formulae for each of them (Equation 1 in the case of cepstrum). This is obviously not the case when computing the objective distance on phone duration.

These objective measures are computed for all the vowels of the synthesis set of the database. The mean MCD, RMSE_lf0 and RMSE_dur, together with their 95% confidence intervals, are shown in Table 3 for each DoA. We observe in this objective evaluation that the MCD increases from HPR to HPO speech, while RMSE_lf0 and RMSE_dur decrease with the DoA. Considering that HPR speech is characterized by longer phone durations and HPO speech by shorter ones (Section 3.2.3):

- modeling the HPR speech cepstrum seems easier, as more speech data are indirectly available to estimate reliably the corresponding models compared to the NEU style (which can be seen with the MCD). On the other hand, modeling HPO speech cepstrum seems more difficult, as less speech data are indirectly available. Note that 1 dB is usually accepted as the difference limen for spectral transparency [Paliwal and Atal (1993)];

- each HPR phone therefore contains a higher number of frames, and errors induced at the frame level by the HMM-based modeling could thus have more impact on the synthesized speech quality, while each HPO phone contains fewer frames, leading to the opposite conclusion. This explains the RMSE_lf0 and RMSE_dur results in Table 3.

Results	HPR	NEU	HPO
Mean MCD ± CI	5.9 ± 0.1	6.3 ± 0.2	6.9 ± 0.1
RMSE_lf0 ± CI	213.1 ± 31.1	170 ± 23.5	112.3 ± 14.3
RMSE_dur ± CI	9 ± 0.6	6.8 ± 0.6	4.6 ± 0.4

Table 3: Objective evaluation results: average MCD [dB], RMSE_lf0 [cent] and RMSE_dur [number of frames] with their 95% confidence intervals (CI) for each DoA.

For comparison purpose, similar quantitative results were observed for speaker dependent model [Yamagishi and Kobayashi (2007)], despite some differences in the training process and in the language used for training the models (Japanese). Differences between the results reported in [Yamagishi and Kobayashi (2007)] and ours are rather minor, as we get slightly worse values for the MCD, slightly better performance for RMSE_lf0 and comparable values for RMSE_dur.

It is also worth noticing that the vocalic space reduces from HPR (0.299 kHz^2) to NEU (0.201 kHz^2) to HPO speech (0.063 kHz^2). These numbers obtained on synthesized speech are to compare to those of Section 3.1.1 which were carried out on the original recordings. Interestingly, a high similarity can be underlined which confirms a good reproduction of DoA changes by HMMs.

4.1.3. Subjective Evaluation

A subjective evaluation has then been performed in order to confirm the results of the objective test. For this evaluation, participants were asked to listen to three sentences: A, the original sentence; B, the copy-synthesis version of the original sentence using the DSM vocoder; C, the sentence synthesized using DSM and whose parameters are generated by the statistical models trained with HTS. Participants were given a 9-point discrete scale and asked to score the distance, in terms of overall speech quality, of B with regards to both A and C. In other words, this score was allowed to vary from 0 (i.e. B has the same quality as A) to 9 (i.e. B has the same quality as C).

The reason for using such a self-designed scale is to ease the listener's ability to evaluate the relative speech synthesis quality. Indeed, evaluating the perceptual position of B between the lower boundary A and the upper one C is more coherent than estimating its position knowing only one of these two boundaries. The latter case brings the problem of the extent until which the score could be set by the listener, and the problem of inter-listener variability concerning evaluations.

The passage from A to C accounts for two possible sources of degradation: vocoding (from A to B) and HMM-based statistical processing (from B to C). Since we can assume that the vocoding effect is almost the same for each DoA, the distance of B with regards to A and C is informative about the effectiveness of the statistical process. Indeed, the lower the score, the more B is close to A than it is from C, and consequently the more dominant is the statistical process among the degradation sources. In conclusion, the higher the score, the better the steps of HMM modeling and generation have been performed.

The test consists of 15 triplets: 5 sentences per DoA randomly chosen amongst the synthesis set, 3 DoA, 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 familiarizing with the scale. During the test, he was allowed to listen to the triplet of sentences as many times as wanted (participants were nonetheless advised to listen to A and C before listening to B, in order to know the boundaries of the scale). However they were not allowed to come back to previous sentences after validating their decisions.

26 people, mainly naive listeners, participated to this evaluation. The mean scores for each DoA, on the 9-point scale, are shown in Figure 5. It is observed that the more speech is articulated, the higher the score and therefore the lower the degradation due to the HMM process. It is worth noting that these results corroborate the conclusions of the objective evaluation. The formant trajectories are enhanced in HPR speech and are less marked in HPO speech, compared to the standard NEU style. Due to the intrinsic statistical modeling by HMMs, these trajectories are (over) smoothed, loosing the actual finest information characterizing the DoA. Among others, this is particularly true for HPO speech.

4.2. Speaking Style Adaptation

One way to perform HMM-based speech synthesis is to train a full data model using a database containing specific data, as in Section 4.1 (in this case, the database contains speech sentences pronounced with different DoA). Compared to unit-selection speech synthesis, HMM-based speech synthesis has many advantages, mainly related to its inherent flexibility due to the statistical modeling process [Zen et al. (2009)]. For example, voice adaptation techniques can be applied to change voice characteristics and prosodic

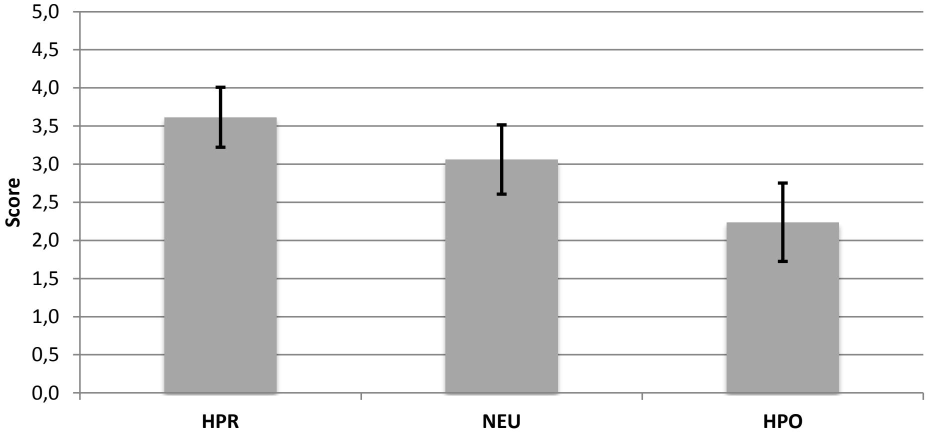

Figure 5: Subjective evaluation results: overall speech quality of the full data models decreases with the DoA (mean score with its 95% confidence interval).

features of synthetic speech [Yamagishi et al. (2009)]. Yamagishi proposed in [Yamagishi and Kobayashi (2007)] the adaptation of a specific model, called average-voice model, to a specific target speaker. The average-voice model is computed once for all over a database containing many different speakers. This technique allows to provide high quality speech synthesis using a limited amount of adaptation data [Yamagishi (2006)].

In this section, we focus on the adaptation of a specific source speaker, the NEU full data model trained in Section 4.1, such that the system is able to generate HPO and HPR speech.

4.2.1. Method

This NEU full data HMM-based speech synthesizer was adapted using the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [Digalakis et al. (1995)] [Gales (1998)] in the framework of Hidden Semi Markov Model (HSMM) [Ferguson (1980)] with HPO and HPR speech data in order to produce respectively a HPO and HPR HMM-based synthesizer. The linearly-transformed models were further optimized using MAP adaptation [Yamagishi et al. (2009)].

In HSMM-based speech synthesis [Zen et al. (2007)], state duration distributions are modeled explicitly, allowing in this way a better representation of the temporal structure of human speech. HSMM has also the advantage of incorporating state duration models explicitly in the expectation step of the Expectation-Maximization (EM) algorithm. Finally, HSMM is more convenient during the adaptation process to simultaneously transform both state output and state duration distributions.

MLLR adaptation is the most popular linear regression adaptation technique. The mean vectors and covariance matrices of state output distributions of the target speaker's model are obtained by linearly transforming the mean vectors and covariance matrices of state output distributions of the source speaker's model [Yamagishi and Kobayashi (2007)]. The same idea holds for CMLLR. While MLLR is a model adaptation technique, CMLLR is a feature adaptation technique. In a model adaptation technique, a set of linear transformations is estimated to shift the means and alter the covariances in the source speaker's model so that each state in the HMM system is more likely to generate the adaptation data. In a feature adaptation technique, a set of linear transformations is estimated to modify the feature vectors in the source speaker's model so that each state in the HMM system is more likely to generate the adaptation data.

The implementation of our synthesizers is summarized in Figure 4. Since the two synthesizers implemented in this section are created by adapting the NEU full data model using HPO and HPR data, they will be referred to as *adapted models* in the following of this work.

The efficiency of the adaptation process will be now assessed through both an objective and a subjective evaluation on the synthesis set of the database, composed of sentences which were neither part of the training set nor of the adaptation set.

4.2.2. Objective Evaluation

The goal of this objective evaluation is to assess the quality of the adapted synthesized speech when the number of adaptation sentences increases. For this, we use the measures introduced in Section 4.1.2, namely the average MCD, the RMSE_lf0 and the RMSE_dur. As an illustration, Figure 6 presents the average MCD, computed for all the vowels of the synthesis set, between the adapted and the full data models.

Figure 6 clearly shows that the MCD decreases when more speech data are used for adaptation. The distance between the HPR full data and adapted models is bigger than the gap between the HPO full data and adapted models, which could be explained by the adaptation process itself. On one hand, the HPR speech spectrum is richer, more variable, complex and enhanced, compared to the NEU style. On the other hand, HPO speech spectrum is smoother and more flat than the NEU speech one. This difference could ex-

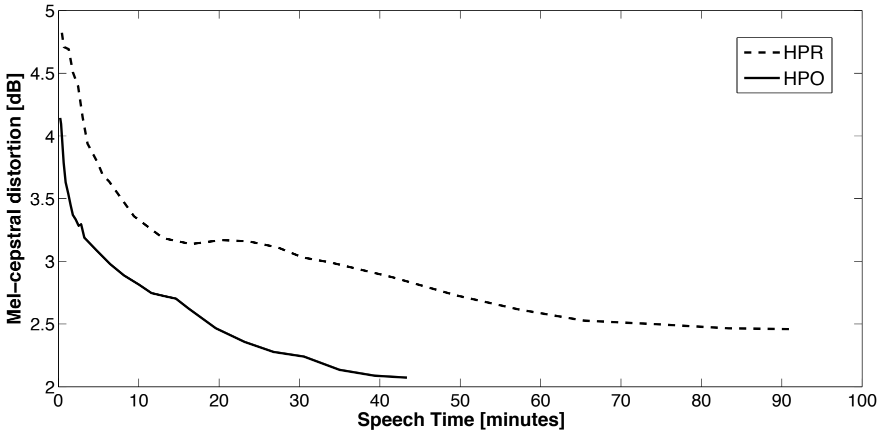

Figure 6: Objective evaluation - Average mel-cepstral distortion [dB] computed between the adapted and the full data models.

plain why the HPR spectrum is harder to adapt from the NEU style (leading to a higher MCD) than the HPO spectrum. Note that the results in Figure 6 were obtained using up to 1220 adaptation sentences for both HPO and HPR speech. Nonetheless, since the speaking rate in HPR speech is known to be much slower than in HPO speech (almost the double - see Section 3.2.4), this explains why these curves do not cover the same total adaptation duration.

We also observed a decrease of RMSE_lf0 and RMSE_dur when the amount of speech data available for adaptation increases. They both were found higher for HPR speech than for HPO speech. However, while the MCD is continuously decreasing when more speech data are used for adaptation, it was shown that RMSE_lf0 and RMSE_dur decrease until around 7 minutes of HPO speech or 13 minutes of HPR speech and saturate to specific values when more speech data are used (the interested reader is referred to [Picart et al. (2011a)] for more details). It can be noted from Figure 6 that around 7 minutes of HPO speech or 13 minutes of HPR speech are needed to adapt cepstra correctly, while it was shown in [Picart et al. (2011a)] that around 3 minutes of HPO speech or 7 minutes of HPR speech are sufficient to adapt F0 and phone duration with a good quality.

Figure 6 also shows some imperfections of the adaptation process based on HMM. Indeed, the curves are saturating towards non-zero values. Slight audible differences could be heard between the HPO or HPR full data models and the models adapted from the NEU full data model using the entire HPO or HPR training set. However, informal listening tests showed that these slight differences cannot be said to give worse or better speech synthesis results. As already stated, 1 dB is usually accepted as the difference limen for spectral transparency [Paliwal and Atal (1993)].

For comparison purpose, the same kind of trends were observed for inter-speaker voice adaptation [Yamagishi and Kobayashi (2007)], despite some differences in the training process and in the number of training and adaptation data.

4.2.3. Subjective Evaluation

A Comparison Category Rating (CCR) evaluation is now performed in order to confirm the conclusions of the objective test. 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 models using 10, 20, 50, 100 or 1220 sentences. 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 the two versions are found to be slightly similar. Listeners were asked to score the overall speech quality of B compared to A. The higher the CCR score, the more efficient the adaptation process. Unlike the objective evaluation, there is no need here to have a one-to-one correspondence between the target and the estimated frames. Therefore audio examples used for this evaluation were entirely generated (i.e. cepstrum, F0 and phone duration) by the full data and adapted HMM-based speech synthesizers.

The test consists of 30 pairwise comparisons. The same experimental protocol as in Section 4.1.3 was applied. 26 people, mainly naive listeners, participated to this evaluation. Figure 7 displays the mean CCR scores for both DoA. The same kind of tendency as in the objective evaluation can be seen, i.e. HTS is able to produce better adapted HPO speech than adapted HPR speech. As expected, we also see that the speech synthesis quality of the adapted models increases with the number of adaptation sentences, independently of the DoA. Nonetheless, a reasonably high-quality HMM-based speech synthesis can be achieved for both DoA with around 100 HPO or HPR adaptation sentences. It can be indeed seen from Figure 7 that this corresponds to CCR scores around 3.5, which means that the adapted voice, compared to the full data model, is perceived to have a quality between “slightly similar” and “similar”.

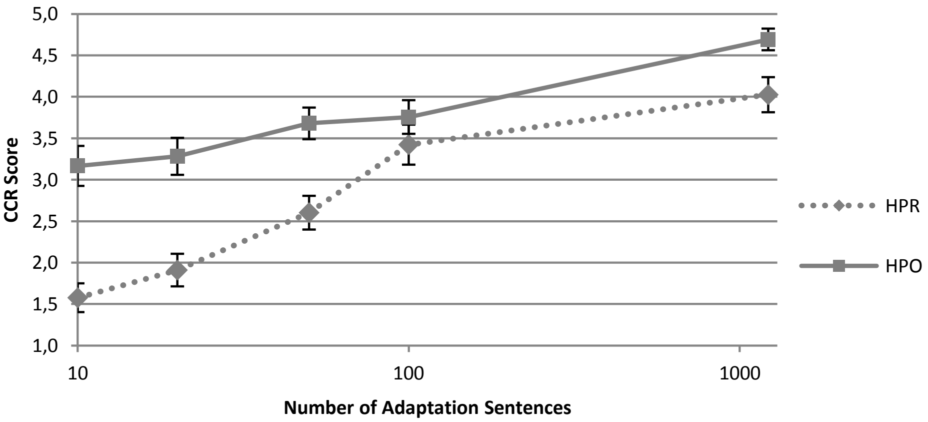

Figure 7: Subjective evaluation of the adapted models - Effect of the number of adaptation sentences on CCR scores (mean scores with their 95% confidence intervals).

4.3. Continuous Control of the Articulation Degree (Interpolation and Extrapolation)

This section is devoted to the implementation and quality assessment of a continuous control of the DoA in HMM-based speech synthesis, in order to continuously and smoothly change the DoA of the NEU voice towards and possibly beyond our adapted HPO or HPR voices.

Thanks to both the statistical and parametric representation used in HMM-based speech synthesis, interpolation between the speaking styles is possible. Speaker interpolation is performed in [Yoshimura et al. (2000)] by interpolating HMM parameters amongst 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. Three main methods for modeling and interpolating between speaking styles have been proposed: style-dependent modeling and style mixed modeling [Yamagishi et al. (2003)]; model interpolation technique [Yoshimura et al. (2000)]; MLLR-based model adaptation technique [Tamura et al. (2001)]. In this latter study, the speaking styles they considered are various emotions, while they refer in our case to the DoA. Dialect interpolation has been performed in [Pucher et al. (2010)] using dialect-dependent and dialect-independent modelings. [Kazumi et al. (2010)] suggested factor-analyzed voice models for creating various voice characteristics in HMM-based speech synthesis. Recently, a computational model of human speech production to manage phonetic contrast along the “H and H” continuum has been proposed and implemented in [Nicolao et al. (2012)], allowing speaking style modification in HMM-based speech synthesis according to the external acoustic conditions.

4.3.1. Method

Our implementation for a continuous control of the DoA makes use of 3 models: *i)* the NEU full data model; *ii)* the adapted HPO model; *iii)* the adapted HPR model, as it is illustrated in Figure 4. Both adapted models were obtained using the entire training HPO and HPR sets (1220 sentences) in order to obtain the finest quality for model interpolation and extrapolation and consequently for the resulting delivered speech synthesis.

Because decision trees of the NEU full data model are not modified during the adaptation process, there is a one-to-one correspondence between the probability density functions (i.e. the leaf nodes of the decision trees) of the NEU full data model and the adapted HPO or HPR models. Therefore the continuous control of the DoA is achieved by linearly interpolating or 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 are available to evaluate objectively the quality of interpolation and extrapolation, only two subjective tests are conducted. The way listeners perceive the interpolation and extrapolation of the DoA is first assessed in Section 4.3.2. This evaluation is then complemented with a Comparative Mean Opinion Score (CMOS) test in Section 4.3.3, to assess the quality of this interpolation and extrapolation.

4.3.2. Perception of the degree of articulation

For this evaluation, listeners were asked to listen to four sentences: the three reference sentences A (HPO), B (NEU) and C (HPR) 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, participants were asked to tell where X should be located on that scale, X being different from A, B or C.

The test consisted of 10 quadruplets. Five sentences per DoA were randomly chosen amongst the synthesis set of the database. 34 people, mainly naive listeners, participated to this evaluation, under the same listening conditions as in Section 4.1.3.

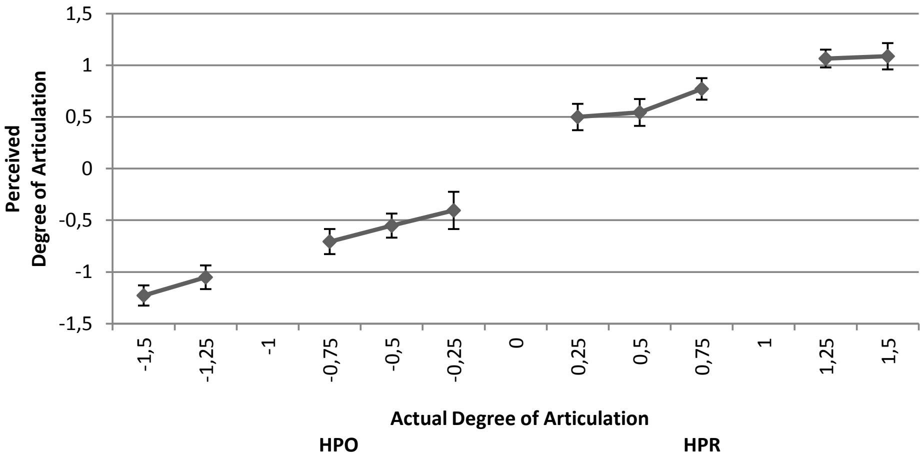

Figure 8: Subjective evaluation - Perceived interpolation and extrapolation ratio as a function of the actual interpolation and extrapolation ratio, together with its 95% confidence interval.

Figure 8 displays the evolution of the average perceived interpolation and extrapolation ratio, as a function of the actual ratio which is applied.

A good linear correspondence is achieved between the perceived and the reference DoA. As expected, this graph is monotonically increasing, showing that listeners were able to perceive and recognize the continuous control of the DoA. Our interpolation and extrapolation method thus proved to be efficient by providing realistic DoA. However, due to the constraints imposed on the discrete scale, i.e. 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 DoA. This bias leads the results to suffer from border effects. Indeed, as participants do not know in advance the maximum variability during the test, they tend to naturally keep out of the border values composing the scale. Extending this scale one point further on both side of a discrete scale, or the use of a continuous scale also extended beyond the range of usual values, should give more accurate results.

4.3.3. Quality of the degree of articulation

In a second subjective test, participants were asked to score the overall speech quality of X versus B (the NEU synthesis), leaving aside the difference in DoA between X and B. For this, we used a CMOS test in order to assess the quality of the interpolated and 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.

Table 4: *Subjective evaluation (CMOS test) - Perceived synthesis quality of the test sentence X vs. the NEU sentence B (CMOS scores with their 95% confidence intervals).*

HPR		HPO	
DoA	Quality	DoA	Quality
0.25	0.03 ± 0.40	- 0.25	- 0.29 ± 0.41
0.5	-0.09 ± 0.46	- 0.5	- 0.47 ± 0.37
0.75	0.15 ± 0.41	- 0.75	- 0.53 ± 0.37
1.25	-0.65 ± 0.49	- 1.25	- 0.35 ± 0.57
1.5	-1.06 ± 0.50	- 1.5	- 0.79 ± 0.57

Table 4 presents the averaged CMOS scores of the perceived synthesis quality for each DoA. The methods proposed in this work provides a high-quality rendering of the DoA. It can be observed that interpolated HPR speech (with a DoA between 0 and 1) seems to have about the same quality as NEU speech, while a slight degradation is observed for all other DoA (on the CMOS scale, a score of -1 means "slightly worse"). Similarly to Section 4.1, HTS provides a better rendering of HPR speech, compared to the HPO speech case. Note also the large size of the 95% confidence intervals for each DoA, and mainly when extrapolating. This could be explained by the difficulty to compare speech quality alone, leaving aside the fact that the DoA of X and B could be different.

5. Perceptual Considerations of Hypo and Hyperarticulated Speech

This section is devoted to the effects leading to and induced by the perception of the DoA. As already mentioned in the introduction, increasing the intelligibility of a synthesizer performing in adverse conditions has a lot of daily life applications. For this, [Erro et al. (2012)] proposed to improve the intelligibility of speech by manipulating the parameters (spectral slope and amplification of low-energy parts of the signal) of an harmonic speech model. Five energy reallocation strategies to increase speech intelligibility in noisy conditions are compared in [Tang and Cooke (2010)]. Regarding the evaluation, the work presented in [Valentini-Botinhao et al. (2011)] interestingly investigated several objective measures to predict the intelligibility of synthetic speech.

The perceptual prevalence of phonetic, prosodic and filter information has been studied in [Picart et al. (2011b)]. The internal mechanisms leading to the perception of each DoA by listeners have been compared and quantified: impact of cepstral adaptation, of prosody, of phonetic transcription and of the complete adaptation technique. All these effects outperformed the baseline system, in which a straightforward phone-independent constant ratio was applied to pitch and phone durations to sound like real HPO and HPR speech. It was shown that adapting the cepstrum has a higher impact on the rendering of the DoA than adapting the phonetic transcriptions. Moreover, adapting prosody alone, without cepstrum adaptation, highly degrades the perception of the DoA. We finally highlighted the importance of having a Natural Language Processor able to create automatically realistic HPO and HPR transcriptions.

As a complement to this latter study, the following sections focus on intelligibility (Section 5.2) and multi-dimension (Section 5.3) assessment of the speech synthesizers described in Section 5.1.

5.1. Method

Five HMM-based speech synthesizers are implemented following the same procedure as in Section 4.3. As a reminder, the NEU full data model was adapted using the entire HPO and HPR training sets, in order to remove the effect of the number of adaptation sentences from our results. The five synthesizers are created using interpolation ratios ranging from -1 (HPO) to +1 (HPR), including 0 (NEU), with a 0.5 step: -0.5 and +0.5 correspond to models right between the NEU full data model and respectively, the adapted HPO model, or the adapted HPR model. The intelligibility of these five synthesizers (-1, -0.5, 0, +0.5, +1) will be studied in Section 5.2, while a general assessment will be performed on the three major synthesizers (-1, 0, +1) in Section 5.3.

5.2. 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 listeners cannot determine a word in the sentence by the meaning of the whole utterance or the context within the sentence.

5.2.1. Building the SUS Corpus

The same corpus as the one built in [de Mareüil et al. (2006)] 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 containing 4 target words (nouns, verbs or adjectives, here written with a capital initial letter):

- 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 4 originally proposed by [Benoît (1990)] was not kept, because it only contained 3 target words instead of 4. For more details about the generation of this corpus, the reader is referred to [de Mareüil et al. (2006)].

5.2.2. Procedure

Nineteen listeners, mainly naive, 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, they 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 5.1. Two types of degradation were then applied to the synthesized SUS: additive noise and reverberation.

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[2] database, and was added

[2] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

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:

$$x(n) = h(n) * s(n), \tag{2}$$

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 with two levels of reverberation (T_{60} of 100 and 300ms) was simulated using the source-image method [Allen and Berkley (1979)], and the simulated impulse responses convolved with original speech signals.

The word level recognition accuracy is used as performance metric for the SUS test. In order to cope with orthographic mistakes, this accuracy was computed by counting manually the number of erroneous phonemes for each word written by the listeners, in comparison with the correct word. The same procedure was also applied for the accuracies at the sentence level. However, they are not displayed on Figure 9 for the sake of conciseness, but can be found in [Picart et al. (2012)]. A strong correlation was noted between the recognition accuracy at the sentence and word levels.

5.2.3. Results

The mean recognition accuracies at the word level (for each DoA, for each type and level of perturbation) are shown in Figure 9. The higher the score, the better the synthesizer intelligibility as it leads to a higher word recognition.

Interestingly, it is observed that accuracy generally increases with DoA. For example, in the strongest reverberation, the word recognition rate increases from around 48% for HPO speech, to 83% in HPR (i.e. an absolute gain of 35%). It is also worth noting that in the presence of car noise, there is no need to over-articulate: using values of 0.5 or 1 for the DoA leads to almost exactly the same intelligibility performance. This conclusion however does not hold in a reverberant environment. Comparing the effect of the perturbation on the message understandability, it turns out that the most

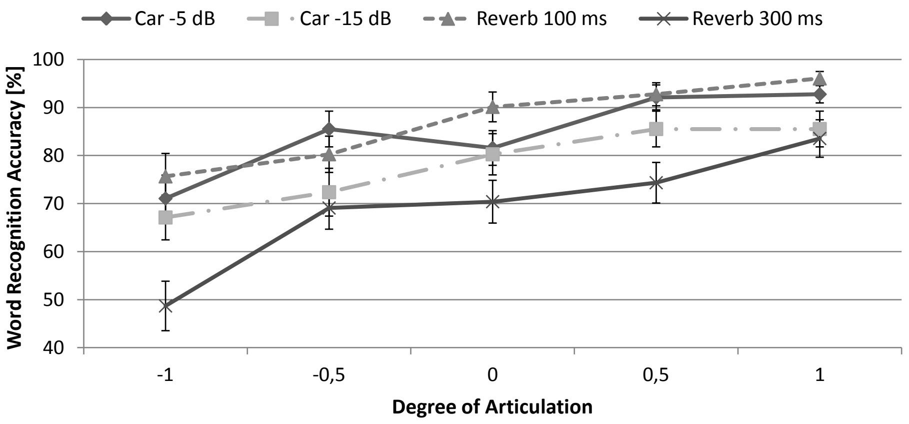

Figure 9: SUS Test - Mean word recognition accuracies [%], together with their 95% confidence intervals.

reverberant condition clearly leads to the highest degradation. In HPR, increasing the level of noise from -5dB to -15dB SNR results in a reduction of the word recognition rate of around 7%. Finally, it is noticed that, on average, the weakest reverberation is the less adverse condition, with recognition rates ranging from 75% to 96% when increasing the DoA. These latter results are curiously observed to be about 9% better than in a car noise with -15dB SNR, whatever the DoA.

5.3. Absolute Category Rating Test

Finally, an Absolute Category Rating (ACR) test was conducted in order to assess several dimensions of the generated speech. As in [de Mareüil et al. (2006)], the Mean Opinion Score (MOS) was complemented with six other categories: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation.

5.3.1. Procedure

Seventeen listeners, mainly naive, 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 [de Mareüil et al. (2006)]). 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 DoA (NEU, HPO or HPR). We used the same listening protocol as in Section 4.1.3.

5.3.2. Results

Results are shown in Figure 10. In all cases, original speech is preferred to synthetic speech. The MOS test shows that original NEU speech is preferred to HPO and HPR speech, while synthetic NEU and HPR speech are almost equivalent, leaving synthetic HPO speech slightly below. The comprehension test points out that NEU and HPR speech are clearly more understandable than HPO speech, both on the original and synthetic side. Differences of comprehension between original and synthesized speech are interestingly rather weak. The pleasantness test indicates a preference of the listeners for original NEU speech, followed by HPR and HPO 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 time, 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 HPO speech has a higher score than the others. This is due to the fact that HPO speech is characterized by a lower number of breaks and glottal stops, shorter phone durations and higher speech rate (as proven in Section 3.2). All these effects lead to an impression of fluidity in speech, while the opposite tendency is observed in HPR 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. Albeit NEU and HPR speech are perceived equivalently in this ACR test from the comprehension and pronunciation points of view, the SUS test proved that HPR speech was much more intelligible than NEU speech in adverse environments.

6. Conclusions

This paper focused on the analysis and synthesis of hypo (HPO) and hyperarticulated (HPR) speech, compared to neutral (NEU) speech. Integrating a continuous variable degree of articulation (DoA) within HMM-

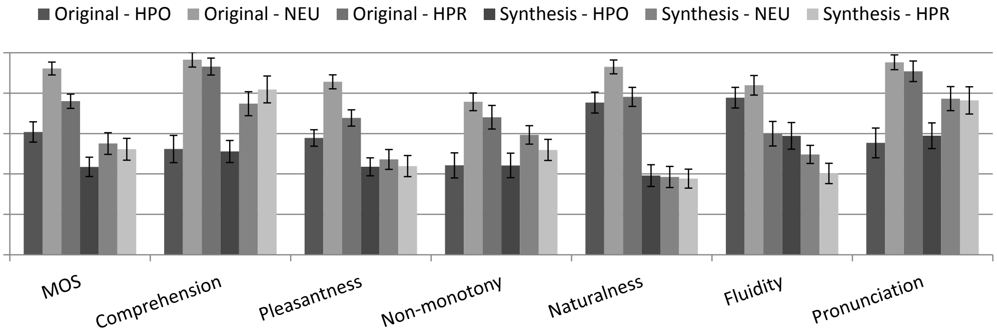

Figure 10: ACR Test - Mean scores together with their 95% confidence intervals.

based speech synthesis is of interest in several applications: expressive voice conversion in embedded systems or for video games, reading speed control for visually impaired people, improving intelligibility performance in adverse environments (e.g. GPS voice inside a moving car, train information in stations), etc. This is also necessary in a more realistic system able to mimic more accurately humans who constantly adapt their speaking style to the communication context.

For this, the paper was divided into three main parts. In the first one, we led a study on the speech modifications occuring when the speaker varies his DoA. At the acoustic level, it was shown that both the vocal tract and glottal contributions are affected. More precisely, an increase of articulation is significantly reflected by an augmentation of the vocalic space in the $F1$-$F2$ plane, by higher $F0$ values, by a stronger harmonicity in speech and by a glottal flow containing more energy in the high frequencies. At the phonetic level, the main variations concern glottal stops, breaks and the phoneme Schwa /@/. Finally, although the speaking rate significantly increases when the DoA decreases, it turns out that the proportion between speech and pausing periods remains constant.

The second part of the paper aimed at developing a HMM-based speech synthesis system incorporating a continuous tuning of the DoA. This goal was subdivided into three tasks: *i)* building a HMM-based synthesizer for each DoA using the full specific datasets; *ii)* for HPO and HPR speech, being able to create a HMM-based synthesizer by adaptation of the NEU synthesizer and using a limited amount of data; *iii)* being able to continuously control the DoA by interpolating and extrapolating existing models. Both objective and subjective tests were used to validate each of these three tasks. Our conclusions showed that: *i)* HPR speech is synthesized with a better quality; *ii)* about 7 minutes of HPO or 13 minutes of HPR speech are required to adapt correctly cepstral features, while only half of it can be used for pitch and duration adaptation; *iii)* the continuous modification of articulatory efforts is correctly perceived by listeners, while keeping an overall quality comparable to what is produced by the NEU synthesizer.

In the third and last part, we have performed a comprehensive perceptual evaluation of the resulting flexible speech synthesizer. First, a Semantically Unpredictable Sentences (SUS) test revealed that playing on the articulation significantly improves the intelligibility of the synthesizer in adverse environments (both noisy and reverberant conditions). Secondly, an Absolute Category Rating (ACR) test was used to assess the synthesizer through various voice dimensions. Although a loss is noticed between natural and synthesized speech regarding its naturalness and segmental quality, several perceptual features like comprehension, non-monotony and pronunciation are relatively well preserved after statistical and parametric modeling.

The study described in this paper focused on the variations of the DoA produced by a French male speaker. Nonetheless, the approach we adopted and the methods we have developed can be transposed to a variety of speaking styles in various languages. In this way, our ongoing research activities encompass the possibility of controlling the DoA for any French speaker (male or female) for whom no recordings of HPR or HPO speech are available. Once this will be done, our goal will be to transpose this flexibility to other languages, which will probably raise some phonetic issues. Finally, our last target will be the application of these methods to other types of expressivity in speech (e.g. emotional speech with happy and sad data), or even to modalities other than speech (e.g. expressive walk or singing voice synthesis).

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).

References

Adda-Decker, M., de Mareüil, P.B., Lamel, L., 1999. Pronunciation variants in french : schwa & liaison, in: 14th International Conference on Phonetic

Science (ICPhS), San Francisco.

Allen, J., Berkley, D., 1979. Image method for efficiently simulating small-room acoustics. Journal of the Acoustical Society of America (JASA) 65, 943–950.

Anumanchipalli, G.K., Muthukumar, P.K., Nallasamy, U., Parlikar, A., Black, A.W., Langner, B., 2010. Improving speech synthesis for noisy environments, in: Speech Synthesis Workshop 7 (SSW7), Kyoto, Japan. pp. 154–159.

Baker, R.E., Bradlow, A.R., 2009. Variability in word duration as a function of probability, speech style, and prosody. Language and Speech 52, 391–413.

Beller, G., 2007. Influence de l'expressivité sur le degré d'articulation, in: RJC Parole, France.

Beller, G., 2009. Analyse et Modèle Génératif de l'Expressivité - Application à la Parole et à l'Interprétation Musicale. Ph.D. thesis. Université Paris VI - Pierre et Marie Curie, IRCAM.

Beller, G., Hueber, T., Schwarz, D., Rodet, X., 2006. Speech rates in french expressive speech, in: Third International Conference on Speech Prosody, Dresden, Germany.

Beller, G., Obin, N., Rodet, X., 2008. Articulation degree as a prosodic dimension of expressive speech, in: Fourth International Conference on Speech Prosody, Campinas, Brazil.

Benoît, C., 1990. An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity. Speech Communication 9, 293–304.

Bonardo, D., Zovato, E., 2007. Speech synthesis enhancement in noisy environments, in: Interspeech, pp. 2853–2856.

Borroff, M.L., 2007. A landmark underspecification account of the patterning of glottal stop. Ph.D. thesis. Stony Brook University, New York.

Bozkurt, B., Dutoit, T., 2003. Mixed-phase speech modeling and formant estimation, using differential phase spectrums, in: VOQUAL, pp. 21–24.

Browman, C.P., Goldstein, L., 1994. “targetless” schwa: an articulatory analysis. Laboratory Phonology II: Gesture, Segment, Prosody 4, 194–219.

Cerňak, M., 2006. Unit selection speech synthesis in noise, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France.

Childers, D.G., Lee, C.K., 1991. Vocal quality factors: Analysis, synthesis, and perception. Journal of the Acoustical Society of America (JASA) 90, 2394–2410.

Cooke, M., King, S., Kleijn, B., Stylianou, Y. (Eds.), 2012. The Listening Talker - An interdisciplinary workshop on natural and synthetic modification of speech in response to listening conditions, Edinburgh, Scotland.

D’Alessandro, C., 2006. Voice source parameters and prosodic analysis. Walter de Gruyter, Sudhoff. pp. 63–87.

Digalakis, V., Rtischev, D., Neumeyer, L., 1995. Speaker adaptation using constrained reestimation of gaussian mixtures. IEEE Transactions on Speech and Audio Processing 3, 357–366.

Drugman, T., Bozkurt, B., Dutoit, T., 2011. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Communication 53, 855–866.

Drugman, T., Bozkurt, B., Dutoit, T., 2012. A comparative study of glottal source estimation techniques. Computer Speech & Language, Elsevier 26, 20–34.

Drugman, T., Dutoit, T., 2010. Glottal-based analysis of the lombard effect, in: Interspeech, Makuhari, Japan. pp. 2610–2613.

Drugman, T., Dutoit, T., 2012. The deterministic plus stochastic model of the residual signal and its applications. IEEE Transactions on Audio, Speech, and Language Processing 20, 968–981.

Erro, D., Stylianou, Y., Navas, E., Hernaez, I., 2012. Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model, in: Interspeech, Portland.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four parameter model of glottal flow, in: STL-QPSR4, pp. 1–13.

Ferguson, J., 1980. Variable duration models for speech, in: Proc. Symp. on the Application of Hidden Markov Models to Text and Speech, pp. 143–179.

Gales, M., 1998. Maximum likelihood linear transformations for hmm-based speech recognition. Computer Speech & Language 12, 75–98.

Garnier, M., Bailly, L., Dohen, M., Welby, P., Loevenbruck, H., 2006a. The lombard effect: a physiological reflex or a controlled intelligibility enhancement?, in: International Seminar on Speech Production (ISSP), Ubatuba, Brazil. pp. 255–262.

Garnier, M., Bailly, L., Welby, M.D.P., Loevenbruck, H., 2006b. An acoustic and articulatory study of lombard speech. global effects at utterance level, in: International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA, USA. pp. 2246–2249.

Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29, 383–406.

Junqua, J., 1993. The lombard reflex and its role on human listeners. Journal of the Acoustical Society of America (JASA) 93, 510–524.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W.D., 2001. Probabilistic relations between words: Evidence from reduction in lexical production. Bybee, Joan and Paul Hopper (eds.). Frequency and the emergence of linguistic structure , 229–254.

Kazumi, K., Nankaku, Y., Tokuda, K., 2010. Factor analyzed voice models for hmm-based speech synthesis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, Texas. pp. 4234–4237.

Keller, E., 2005. The analysis of voice quality in speech processing. Lecture Notes in Computer Science, Springer-Verlag , 54–73.

Klatt, D., Klatt, L., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America (JASA) 87, 820–857.

Langner, B., Black, A.W., 2004. Creating a database of speech in noise for unit selection synthesis, in: Speech Synthesis Workshop 5 (SSW5), Pittsburgh, USA. pp. 229–230.

Langner, B., Black, A.W., 2005. Improving the understandability of speech synthesis by modeling speech in noise, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, Pennsylvania, USA.

Laver, J., 1994. Principles of Phonetics. Cambridge Textbooks in Linguistics, Cambridge University Press, Great Britain.

Lindblom, B., 1983. Economy of Speech Gestures. Spinger-Verlag, New-York.

Lombard, E., 1911. Le signe de l'élévation de la voix. Annales des Maladies de l'Oreille et du Larynx 37, 101–119.

Lu, Y., Cooke, M., 2009. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Communication 51, 1253–1262.

Malfrere, F., Deroo, O., Dutoit, T., Ris, C., 2003. Phonetic alignement : speech-synthesis-based versus viterbi-based. Speech Communication 40, 503–515.

de Mareüil, P.B., d'Alessandro, C., Raake, A., Bailly, G., Garcia, M.N., Morel, M., 2006. A joint intelligibility evaluation of french text-to-speech synthesis systems: the evasy sus/acr campaign, in: 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy. pp. 2034–2037.

Nicolao, M., Latorre, J., Moore, R.K., 2012. C2h: A computational model of h&h-based phonetic contrast in synthetic speech, in: Interspeech, Portland, Oregon, USA.

Nose, T., Tachibana, M., Kobayashi, T., 2009. Hmm-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation. IEICE Transactions on Information and Systems 92, 489–497.

Oviatt, S., Levow, G.A., Moreton, E., MacEachern, M., 1998. Modeling global and focal hyperarticulation during human–computer error resolution. Journal of the Acoustical Society of America (JASA) 104, 3080–3098.

Paliwal, K.K., Atal, B.S., 1993. Efficient vector quantization of lpc parameters at 24 bits/frame. IEEE Transactions on Speech and Audio Processing 1, 3–14.

Pantazis, Y., Stylianou, Y., 2008. Improving the modeling of the noise part in the harmonic plus noise model of speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4609–4612.

Patel, R., Everett, M., Sadikov, E., 2006. Loudmouth: modifying text-to-speech synthesis in noise, in: 8th International ACM SIGACCESS Conference on Computers and Accessibility (Assets), Maltimore, Maryland, USA. pp. 227–228.

Picart, B., Drugman, T., Dutoit, T., 2010. Analysis and synthesis of hypo and hyperarticulated speech, in: Speech Synthesis Workshop 7 (SSW7), Kyoto, Japan. pp. 270–275.

Picart, B., Drugman, T., Dutoit, T., 2011a. Continuous control of the degree of articulation in hmm-based speech synthesis, in: Interspeech, Firenze, Italy. pp. 1797–1800.

Picart, B., Drugman, T., Dutoit, T., 2011b. Perceptual effects of the degree of articulation in hmm-based speech synthesis, in: NOLISP Workshop, Las Palmas, Gran Canaria. pp. 177–182.

Picart, B., Drugman, T., Dutoit, T., 2012. Assessing the intelligibility and quality of hmm-based speech synthesis with a variable degree of articulation, in: The Listening Talker (LISTA) workshop, Edinburgh, Scotland.

Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., Strom, V., 2010. Modeling and interpolation of austrian german and viennese dialect in hmm-based speech synthesis. Speech Communication 52, 164–179.

Raitio, T., Suni, A., Vainio, M., Alku, P., 2011. Analysis of hmm-based lombard speech synthesis, in: Interspeech, Florence, Italy. pp. 2781–2784.

Roekhaut, S., Goldman, J.P., Simon, A.C., 2010. A model for varying speaking style in tts systems, in: Fifth International Conference on Speech Prosody, Chicago, IL, USA.

Sjolander, K., Beskow, J., 2000. Wavesurfer - an open source speech tool, in: Sixth International Conference on Spoken Language Processing (ICSLP), Beijing, China. pp. 464–467.

Södersten, M., Hertegård, S., Hammarberg, B., 1995. Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women. Journal of Voice 9, 182–197.

Stylianou, Y., 2001. Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing 9, 21–29.

Summers, W.V., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M., 1988. Effects of noise on speech production: acoustic and perceptual analyses. Journal of the Acoustical Society of America (JASA) 84, 917–928.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2003. Hmm-based speech synthesis with various speaking styles using model interpolation and adaptation. IEICE Technical Report 103, 37–42.

Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., 2001. Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, USA. pp. 805–808.

Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions, in: Interspeech, Makuhari, Japan.

Tokuda, K., Kobayashi, T., Masuko, T., Imai, S., 1994. Mel-generalized cepstral analysis - a unified approach to speech spectral estimation, in: International Conference on Spoken Language Processing, pp. 1043–1046.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., Zen, H., 2012a. Cepstral analysis based on the glimpse proportion measure for improving the intelligibility of hmm-based synthetic speech in noise, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan. pp. 3997–4000.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2011. Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise?, in: Interspeech.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012b. Mel cepstral coefficient modification based on the glimpse proportion measure for improving the intelligibility of hmm-generated synthetic speech in noise, in: Interspeech, Portland, Oregon, USA.

Wouters, J., 2001. Analysis and Synthesis of Degree of Articulation. Ph.D. thesis. Dept. of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology.

Yamagishi, J., 2006. Average-Voice-Based Speech Synthesis. Ph.D. thesis. Tokyo Institute of Technology.

Yamagishi, J., Kobayashi, T., 2007. Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training. IEICE Transactions Information and Systems 90, 533–543.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., Renals, S., 2009. A robust speaker-adaptive hmm-based text-to-speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing 17, 1208–1230.

Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2003. Modeling of various speaking styles and emotions for hmm-based speech synthesis, in: Proceedings of Eurospeech, Switzerland. pp. 2461–2464.

Yegnanarayana, B., Rajendran, S., Worku, H.S., N., D., 2008. Analysis of glottal stops in speech signals, in: Interspeech, pp. 1481–1484.

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., Kitamura, T., 2000. Speaker interpolation for hmm-based speech synthesis system. Journal of the Acoustic Society of Japan (E) 21, 199–206.

Yuan, J., Liberman, M., Cieri, C., 2006. Towards an integrated understanding of speaking rate in conversation, in: Interspeech, Pittsburgh, PA, USA. pp. 541–544.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Communication 51, 1039–1064.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007. A hidden semi-markov model-based speech synthesis system. IEICE Transactions on Information and Systems 90, 825–834.