C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech

Mauro Nicolao¹, Javier Latorre², Roger K. Moore¹

¹Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK
²Toshiba Research Europe Ltd., Cambridge Research Laboratory, UK
m.nicolao@dcs.shef.ac.uk, javier.latorre@crl.toshiba.co.uk, r.k.moore@dcs.shef.ac.uk

Abstract

This paper presents a computational model of human speech production based on the hypothesis that low energy attractors for a human speech production system can be identified, and that interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be motivated by energetic considerations of phonetic contrast. An HMM-based speech synthesiser along with continuous adaptation of its statistical models was used to implement the model. Two adaptation methods were proposed for vowel and consonant models and their effectiveness was tested by showing that such hypo/hyper-articulation control can manipulate successfully the intelligibility of synthetic speech in noise. Objective evaluations with the ANSI Speech Intelligibility Index indicate that intelligibility in various types of noise is effectively controlled. In particular, in the hyper-articulation transforms, the improvement with respect to unadapted speech is above 25 %.

Index Terms: reactive speech synthesis, hypo/hyper-articulated speech, intelligibility enhancement.

1. Introduction

The observation that human talkers adapt their speech according to the listening situation was established almost a century ago by Lombard [1]. According to theories such as Lindblom’s H&H (hypo-hyper) theory [2] of speech production, such modifications are caused by the need to transfer information from the talker to the listener. Humans make continuous adjustments while they are speaking, continuously assessing the effectiveness of their modifications. In Levelt’s Perceptual Loop theory [3], this adaptation is described as an inner talker process, driven by a perceptual loop which constantly monitors the outcome to ensure the success of the communication process. To obtain such a goal, Moore [4] affirmed that human speakers must be aware of listener’s needs, and that speakers react to the context in which they are speaking, using an internal model of the listener.

State-of-the-art speech synthesis systems still exhibit a rather limited range of speaking styles as well as an inability to adapt to the listening conditions in which they operate. Therefore, Moore introduced the new idea of a reactive speech synthesizer that would monitor the effect of its output and modify its characteristics in order to maximize its communicative intentions. Moore suggested that systems should talk clearly and start to address behaviours exhibited by human talkers such as Lindblom’s hypo/hyper-articulated speech. Such behaviour is intrinsically related to the effort used by talkers during speech production and, in synthesised speech, it can be modelled acting on the energy distribution and organisation [5].

In the experiment reported here, the hypothesis has been evaluated that there are low-energy attractors in the human speech production system and, moreover, that an interpolation/extrapolation along the key dimension of hypo-hyper-articulation can be obtained by controlling the distance to such attractors. Several hypotheses on the nature of these attractors have been tested by modifying either vowel or consonant production in a HMM-based synthesiser with a scalable adaptation of its statistical models.

This experiment is also part of a wider programme of research which aims to create a computational model of human speech production providing continuous adjustment according to environmental conditions. The possibility has thus been investigated for introducing a feedback path into a TTS system such that it could perceive the effectiveness of the communicative process with an objective measurement.

In order to test the degree of control on hypo/hyper-articulation speech, a possible implementation of this model was developed and the intelligibility of the outcome evaluated.

2. The C2H model

Following Moore’s PRESENCE model [4] and Levelt’s theory [3], a computational model (Figure 1) was designed to be as close as possible to a human’s speech production system with the relevant perceptual feedback loop. The model consisted of a speech production system and a negative feedback loop which, respectively, generates the utterances and measures the environment effects on the outcome such that adjustments can be made dynamically according to the results of the analysis. The perceptual feedback consists of an emulation of a listener’s auditory system that measures the environmental state and returns information that is used to control the degree of modification to the speech production. The model allows for several modifications at different levels, even though this experiment is focused on the model adaptation only.

2.1. Implementation of the model

The implementation of this model, named Computational model for H&H theory (C2H), was designed using the latest release of the HMM-based speech synthesiser, HTS[1]. A parameter generation method, the recursive search algorithm, was added to create a reactive speech synthesizer which could handle the adaptations of the statistical models at the frame level.

¹website: http://hts.sp.nitech.ac.jp
3. Actively managing the phonetic contrast

As in [10], the adaptation trained in this experiment did not rely on a data-driven approach. Instead its characteristics were motivated by both articulatory and energetic considerations of phonetic contrast, and by the suggestion that less contrast in the acoustic realisation of phones might give the impression to the listeners of less well articulated speech and of a smaller amount of energy involved in its production [11]. This direct link to the definition of hypo/hyper-articulated speech leads to the notion of low-energy attractors: minimally-contrastive realisations towards which at least two competitive phones tend to converge.

In HMM-based syntheses the acoustic realisation of phones can be changed continuously in any direction in the high-dimensional space defined by their parametric representation with reasonably simple adaptation. Therefore, once identified, a low-energy attractor in the acoustic space defines a specific vector along which it should be possible to move in order to decrease or increase the degree of articulation.

The hypothesis is thus that by manipulating the acoustic distance among phone realisations, it is possible to go from hypo-articulated speech (i.e. by moving towards the attractor) to hyper-articulated (i.e. by moving in the opposite direction away from the attractor) with direct consequences for the intelligibility of the resulting output speech.

3.1. Vowel production control (VPC)

One of the most important consequences of hypo- and hyper-articulated speech has been observed in the characteristic of their vowel spaces. It was reported in both human [12] and synthetic speech (trained on specific corpora) [13] that a reduction happens in a hypo-articulated vowel space while an expansion appears in a hyper-articulated one.

The neutral vowel position defines a specific low-energy configuration in which all other vowels tend to converge when the phonetic contrast among them is perceived to be reduced. Hence, the mid-central vowel schwa can be considered to be the low-energy attractor for the vowels in human speech production, and the dimension of H&H variation is the degree of deviation from that attractor (Figure 2a). Preliminary results of such Vowel Production Control (VPC) were reported in [14].

3.2. Consonant production control (CPC)

An attractor system was also hypothesised for consonants. In this domain, a unique location cannot be identified as in the vowel domain [15], but specific attractors were introduced for different phones. Once identified an acoustically close competitor for every consonant, it can be assumed that every realisation along the dimension identified by the two phones is less contrastive than the original one. In particular, the medium point can be hypothesised to be the Lowest-Contrastive attractive configuration (LC config in Figure 2b) for both phones and the Highest-Contrastive one (HC config in Figure 2b) is achieved by a point in the opposite direction.

Highly-confusable consonant pairs were therefore identified, and transformations were trained to convert one phone into another to allow for Consonant Production Control (CPC). Pair choices can be motivated by several needs: the control of voiced-unvoiced contrast (e.g. [t] vs. [d], etc.) or of confusion in noise (e.g. [t] vs. [p], etc.) as per confusion matrices in [16].

3.3. MLLR transforms

The MLLR transformations were estimated using a corpus of synthetic hypo-articulated speech. This consisted of speech generated using the HTS system with an American English fe-
male voice trained on the ‘CMU-ARCTIC SLT’ corpus and forcing its input control sequences to have only low-energy attractors in them. In the VPC training, all vowels were substituted with a schwa realisation, while in CPC, consonants were changed into their specific competitors. Using decision-tree-based clustering, HTS found the most likely acoustic model according to phonetic and prosodic context for all of the phones, even those unseen in its original training corpus.

Both acoustic- and duration-model adaptations were trained to match the characteristics of the hypo-articulation reference. A set of transformations was obtained which modify the mean vectors in the HMM model descriptions. The covariance vector was not considered. The linear transform can be written as,

$$\mu_i' = A_i \mu_i + b_i \tag{1}$$

where $A_i$ is a $P \times P$ matrix, $b_i$ is a $P \times 1$ bias vector for the $i$-th model, and $P$ is the size of the parametric representation.

As mentioned above, MLLR transformations can be scaled with different strength. Given $\mu_i'$ as per (1), the scaled mean vector, $\mu_i^{(\alpha_i)}$, where $\alpha_i$ is the weighting factor ($\alpha_i \geq 0$) to scale the hypo-articulation transform, is computed as:

$$\mu_i^{(\alpha_i)} = \mu_i + \alpha_i (\mu_i' - \mu_i) = \alpha_i \mu_i' + (1 - \alpha_i) \mu_i \tag{2}$$

The transformation towards hyper-articulated speech is defined as the inverse of the trained transformations (same domain, same directions but opposite sense). It can be demonstrated that, if the weighting factor to the hypo-articulation transform, $\alpha_H$, is defined as $\alpha_H \leq 0$, then (2) can still be applied and, moreover, a continuum of values, in which the formula is defined, is created. From this point onwards, $\alpha_H$ and $\alpha_M$ are addressed as $\alpha$. A discussion about the range of values for $\alpha$ and the effects of (2) when $\alpha \geq 0$ (hypo-) and $\alpha \leq 0$ (hyper-articulation) can be found in the following section.

4. Hypothesis evaluation

The C2H model was used to test the proposed adaptation providing the control of the speech production and the evaluation the outcome. Synthesised speech samples with different-strength MLLR adaptation were produced, and objective measurements were extracted for each generated file and different types of noise. The phonetic analysis was provided by the standard Festival tools1 and the duration control was left to the statistical model and its adaptations.

A set of 200 text sentences from the Blizzard Challenge 20102 evaluation test were used to generate the full-strength direct transformation ($\alpha = \max \alpha$) and full-strength inverse transformation ($\alpha = \min \alpha$) samples. A standard speech synthesis ($\alpha = 0$) of the same set was also provided as reference to compare the degree of modifications. These samples were named respectively fully hypo-articulated (HYO), fully hyper-articulated (HYP), and standard (STD) speech.

The $\alpha$ range needs to be assessed carefully. While in VPC the full-direct transform leads to the low-energy point, for CPC the same point is obtained with roughly a half-intensity transformation. For the inversion is even more complicated because no restriction on the minimum $\alpha$ that can be applied is given, and the risk of an inverse transformation that can produce some unnatural speech phenomena is quite high. Therefore, the boundaries for $\alpha$ were defined by some preliminary tests in which

\footnotesize{\begin{itemize}
  \item 1http://festvox.org/cmu_arctic
  \item 2http://www.cstr.ed.ac.uk/projects/festival/
  \item 3http://www.synsig.org/index.php/Blizzard_Challenge_2010
\end{itemize}}

\textbf{Figure 3:} Modification on the F1-F2 chart caused to VPC adaptation: STD vowel values are shown with black circles, HYO vowels with blue-dashed ones in 3a), and HYP vowels with red-dashed ones in 3b). Formants were extracted with Praat software (\url{http://www.fon.hum.uva.nl/praat/}) and phones are displayed with the “CMU Pronouncing Phoneme Set” (\url{http://www.speech.cs.cmu.edu/cgi-bin/cmudict}).

different values, $\alpha \in [-2, 2]$, were applied and the quality of the outcome was evaluated by, e.g., extracting first and second formant mean values for every vowel. The admissible interval of realistic $\alpha$ values was found to be [-0.8,1] for VPC and [-0.7,0.6], in CPC. Hence, VPC-HYO was defined with $\alpha = 1$ and VPC-HYP with $\alpha = -0.8$, whilst CPC-HYO had $\alpha = 0.6$ and CPC-HYP $\alpha = -0.7$.

A brief analysis of the effect of VPC on formants for such values is displayed in Figure 3. In VPC-HYO the vowel space is effectively reduced and the normal vowels tend to converge to the central part of the plot (see Figure 3a). However, this is not a unique point, and this is confirmed by the observation in human speech production where schwa is not just a centralised vowel but a sound that is assimilated with its phonetic and prosodic context[17]. In VPC-HYP, on the other hand, modifications are less evident (see Figure 3b). This is due to the ‘almost hyper-articulated’ characteristic of the samples used to train the TTS which create an already expanded STD vowel space. Nonetheless, the difference observed in humans between spontaneous and read speech (cfr. Figure 2 in [15]) has a similar tendency.

The main assessment of the adaptation was performed by evaluating control of intelligibility in various noisy environments: real car noise, babble noise recorded in a large meeting room, and real speech of 3-4 people in a meeting room.

In order to normalise the energy of the speech with respect to the noise, the Segmental SNR (SSNR) [8] was computed and, whilst the noise was kept at the same intensity of 70 dB SPL, the speech was amplified to have mean SSNR constant. The SII differences of both HPO and HYP samples with respect to the STD ones are reported in Figure 4 for a single type of noise (car noise) and SSNR ≈ 0 dB. It is worth noting that even if SSNR is constant, the global SNR is not constant.

A complete overview on all the results for each experiment is displayed in Figure 5. Analysing the results, it can be concluded that, on average, in HYO the speech intelligibility was reduced by 25 % and in HYP it was increased by almost the same amount. No significant differences have been observed by applying the control either to vowels (VPC) or to consonants (CPC). The resulting local and overall changes in the utterance intensity and durations played an important role in these results.
5. Conclusion

A computational model to manage phonetic contrast along the H&H continuum (C2H) has been proposed and implemented using a modified HMM-based speech synthesiser with a novel generation algorithm which enables it to react to sudden changes in the external acoustic conditions. The scalable adaptation along the H&H axis used to transform the acoustic model emerged from phonetic-contrast motivation alone. The evaluation with SII showed the desired control of intelligibility in noisy environment, and we expect that future subjective listening tests can confirm these results.

C2H proved to be an effective tool for continuously modifying the quality of synthetic speech and thus confirmed the potential of reactive speech synthesis based on the principle of auditory loop analysis. A good emulation of the human speech recognition system is crucial to identify potential phonetic confusions in a synthesised utterance, hence experiments have started to evaluate the synthesiser’s output in noise using models such as an Automatic Speech Recogniser (ASR).

6. Acknowledgements

The research leading to these results was funded by the EU-FP7 under grant agreement n. 213850 - SCALE.

7. References


Figure 4: Distribution of the SII-value differences (in percentage) between HYO and STD speech (blue-crossed histograms), and between HYP and STD speech (red-dotted histograms) controlled with both adaptation methods at SSNR = 0 dB.

Figure 5: Mean SII-differences (in percentage) between STD speech and the ones adapted with VPC and CPC for all types of noise and SSNR levels (SSNR = -5 dB black, SSNR = 0 dB gray, and SSNR = 5 dB white bars).