REACTIVE SPEECH SYNTHESIS: ACTIVELY MANAGING PHONETIC CONTRAST ALONG AN H&H CONTINUUM

Roger K. Moore and Mauro Nicolao

Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK
r.k.moore@dcs.shef.ac.uk m.nicolao@dcs.shef.ac.uk

ABSTRACT

Notwithstanding the tremendous progress that has taken place in the science and technology of text-to-speech synthesis, state-of-the-art 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. Such behaviour is typical of human speech production and, in 2006, Moore proposed a new type of 'reactive' speech synthesizer that would monitor the effect of its output and modify its characteristics in order to maximize its communicative intentions. This paper presents an investigation into reactive speech synthesis based on the hypothesis that the neutral vowel represents a low energy attractor for a human speech production system, and that interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be motivated by energetic considerations of phonetic contrast. The results indicate that the intelligibility of synthesized speech can be manipulated successfully, thus confirming the potential of reactive speech synthesis.

Keywords: reactive speech synthesis, feedback control, hypo/hyper-articulation.

1. INTRODUCTION

Recent years have witnessed considerable progress in both the science and technology of 'spoken language output' (SLO) [1]. Advances in areas such as linguistic text analysis, natural language generation, supra-segmental and segmental modelling, coupled with the immense growth in available computational resources, have lead to the successful commercialization of a range of high-quality text-to-speech (TTS) systems. As in the field of automatic speech recognition (ASR), TTS has benefited from the introduction of a 'data-driven' paradigm in which large corpora of annotated speech recordings are used, either to provide an inventory of acoustic segments from which selected units are concatenated together to produce the output speech or, as in the most recent research systems, to estimate the parameters of underlying statistical models (using hidden Markov model based approaches) [2].

However, whilst current systems are impressive in comparison to their earlier more hand-crafted counterparts, state-of-the-art TTS systems still exhibit a rather limited range of speaking styles and a general lack of expressiveness [3]. In particular, unlike human talkers, no contemporary TTS system makes dynamic adjustments to its spoken output (e.g. speaking louder or articulating more clearly) in response to the difficulties that may be imposed on its listener(s) by the prevailing acoustic environment or communicative situation.

The observation that human talkers adapt their speech according to the listening situation was established exactly a century ago by Lombard [4] but, until recently, the implications for speech technology have mainly been confined to attempting to compensate for the effect in ASR [5]. In an invited paper on the future of spoken language output published at EUROSPEECH in 2003, Moore listed five challenges that would be required in order for future SLO systems *"to produce believable behaviour that is appropriate to a suitably individualized communicative agent"* [6]. Moore's first challenge suggested that systems should talk 'clearly', and he noted that no contemporary TTS systems had addressed Lindblom's classic 'H&H' (hypo-hyper) behaviour exhibited by human talkers [7]. Moore went on to develop this particular idea further and, in a keynote talk at the 2006 EU IST Conference in Helsinki [8], he proposed a new approach to speech generation that (i) selects its characteristics appropriate to the needs of the listener, (ii) monitors the effect of its own output, and (iii) modifies its behaviour according to its internal model of the listener. He subsequently termed this behaviour '*reactive speech synthesis*', and it became one of the core principles in his more general model of spoken language processing (by mind or machine) [9][10].

2. REACTIVE SPEECH SYNTHESIS

2.1. Synthesis-by-analysis

The general principle of reactive speech synthesis (or '*synthesis-by-analysis*') exploits the ability of negative feedback control processes to monitor and adjust behaviour to achieve an intended perceptual effect [11]. The basic idea is illustrated in Fig. 1.

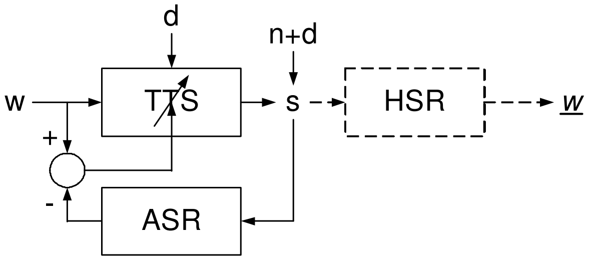

Figure 1: Basic architecture for a reactive speech synthesizer in which words (*w*) are converted to speech (*s*) which is subjected to noise (*n*) and disturbance (*d*). The synthesizer estimates the words perceived by the listener ($\underline{w}$) using a feedback path involving ASR. A control loop compares *w* and $\underline{w}$ and the error signal drives the TTS to alter its output in such a way as to maximize recognition accuracy.

The overall approach can effectively be described as '*synthesis-by-recognition*' (SbR) [9], and it conforms well to contemporary models of human speech production which posit both internal and external control feedback loops for monitoring and modifying behaviour [12][13].

As yet, only a few studies have been reported which actively pursue the SbR approach (e.g. [14]). Establishing a complete feedback loop using both TTS and ASR is a major technical challenge. As a consequence, recent research has focused on collecting and analyzing hypo and hyper speech and applying their characteristics to synthesized speech as a form of 'speaking style' [15].

2.2. Feedback control

One of the key properties of a negative feedback control process is that, in principle, it provides a very efficient mechanism for overcoming *arbitrary* disturbances. In other words, it is *not* necessary to train a model on a range of speaking styles, and then select one according to the prevailing communicative conditions. Rather, the challenge is to determine the appropriate control strategy that would allow synthesized speech to be interpolated and/or extrapolated along the required H&H dimension.

The study reported here is an investigation into the active manipulation of speech synthesis motivated by both articulatory and energetic considerations of phonetic contrast. In particular, it is hypothesized that the neutral vowel [ə] represents a low energy attractor for a human speech production system, and that one dimension of H&H variation which may be of particular interest is the degree of deviation from that attractor. Such an approach is somewhat similar to techniques that have been investigated for speech intelligibility enhancement [16]. The difference here is (i) the motivated use of talker-specific [ə] to anchor the coordinates of the space in which such adjustments are made, and (ii) that the consequences of such adjustments are manifest in both the spectral *and* the temporal domains.

3. EXPERIMENTS

3.1. Experimental setup

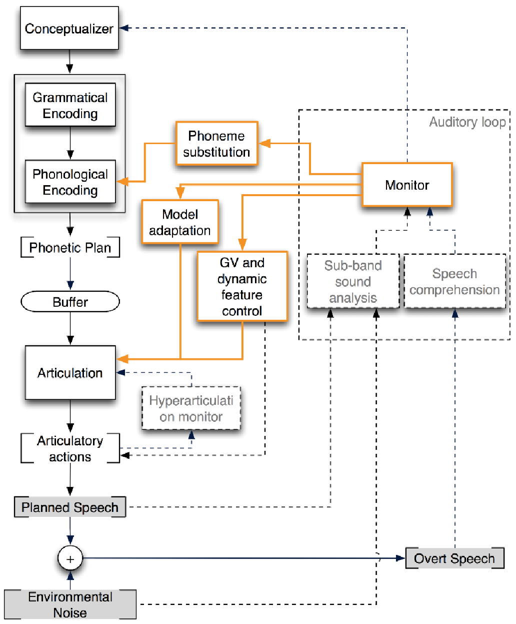

Figure 2: Architecture of the experimental setup used in this study. The 'feedforward' components are to the left, and the 'feedback' components are to the right.

The architecture used to investigate the proposed hypo/hyper-speech transformation is illustrated in Fig. 2. The 'feedforward' components (on the left-hand side of the Figure) comprise the grammatical and phonological encoding from the FESTIVAL TTS system [17] and speech generation using

HMM-based synthesis (HTS). The 'feedback' components (on the right-hand side of the Figure) would, in principle, use an active loop measure of speech intelligibility (e.g. by analyzing the output of an ASR system). In the experiments reported here, it was sufficient to compute the speech intelligibility off-line.

3.2. Transforming the speech

The basic principle that has been investigated is that, whilst it is the case that a given vocalic speech sound can be manipulated in *any* direction in the high-dimensional space defined by its parametric representation (which, in HTS, consists of the parameters of the context-dependent HMMs), the particular location of the neutral schwa vowel [ə] defines a *specific* vector along which it should be possible to produce output with either hypo-articulation (i.e. by moving towards [ə]) or hyper-articulation (i.e. by moving in the opposite direction away from [ə]). The hypothesis is thus that manipulating vowel qualities in this motivated way should have direct consequences for the intelligibility of the resulting output speech.

The required transformation was obtained using a technique that is normally used to adapt HMM-based synthesis to new speakers [18]. A weighted maximum likelihood linear regression (CMLLR) was trained to map normally articulated vowels onto [ə], and the resulting transformation was then applied to newly generated speech. The outcome should be *hypo*-articulated speech. If the hypothesis being put forward in this paper is correct, then applying the inverse transformation should give rise to *hyper*-articulated speech.

Both the forward and the inverse transformations were applied at different strengths, representing different operating points along the derived H&H axis.

3.3. Obtaining the transformation

The CMLLR transformation was estimated using a relatively small corpus of hypo-articulated speech. This consisted of synthesized speech generated by forcing the phone control sequences (around 1100 from the HTS-demo training set) to have only schwa vowels in them. Using decision tree based clustering, HTS found the most likely acoustic model for all of the phones, even those unseen in its original training corpus.

3.4. Test environment

The forward and inverse transformations were applied to a subset of the Blizzard 2010 news corpus [19] (174 sentences) and mixed with real street noise at +5 dB signal-to-noise ratio. The Speech Intelligibility Index (SII) of the resulting synthesized speech was then assessed objectively using the ANSI standard [20].

4. RESULTS

The results of the experiment are presented in Fig. 3. The Figure shows the distribution of SII values across the 174 test utterances for the forward and inverse transformations. As predicted, the forward transformation reduces the measured intelligibility of the synthesized speech. More importantly, the inverse transformation increases the intelligibility (albeit marginally). This confirms that the transformation was operating successfully along an appropriate H&H scale.

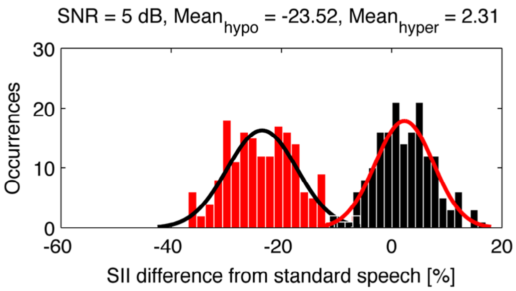

Figure 3: Distribution of SII differences for hyper-articulated (to the right) and hypo-articulated (to the left) speech relative to the untransformed speech.

Clearly the forward transformation to hypo-articulated speech was significantly more effective than the inverse transformation to hyper-articulated speech. Initially it was felt that this represented a flaw in the procedure. However, it was evident that regular synthesized speech has a tendency to be hyper-articulated, and this is especially true for the output from an HMM-based system, which is typically trained on read (i.e. citation form) speech.

In order to confirm this hypothesis, the results were analyzed for a subset of 83 utterances which were measured to have low initial SII scores. The results for this subset are illustrated in Fig. 4. As can be seen, the inverse transformation on the low-SII utterances was considerably more successful, thus confirming the conjecture that, on average, synthesized speech tends to be hyper-articulated.

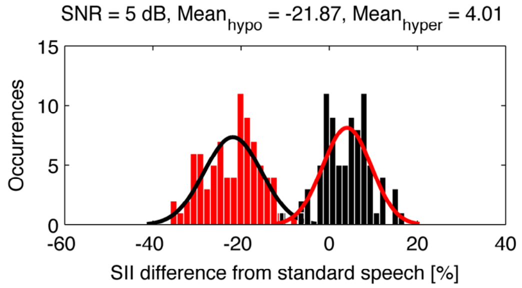

Figure 4: Distribution of SII differences for hyper-articulated (to the right) and hypo-articulated (to the left) speech relative to the untransformed speech for utterances with low SII.

5. SUMMARY & CONCLUSIONS

This paper has reported the results of a study into 'reactive speech synthesis' in which the generation of synthesized speech has been manipulated along an H&H axis based on the use of [ə] as a low energy attractor in the high-dimensional space of phonetic contrasts. The required transformation was obtained using CMLLR (weighted maximum likelihood linear regression), and forward and inverse transformations were applied to a corpus of synthesized utterances mixed with noise. The results confirm that the transformation was operating successfully along an appropriate H&H axis. The intelligibility gains for the inverse transformation towards hyper-articulated speech were marginal, suggesting that synthesized speech is typically already towards the hyper-articulated end of the H&H scale. This was confirmed through an analysis of the results for a subset of the data which revealed significant gains for utterances with low initial intelligibility.

This study has confirmed the potential for Moore's vision of a new approach to speech generation based on 'reactive speech synthesis'. Research continues into methods which would allow a speech synthesizer to select its characteristics appropriate to the needs of the listener, to monitor the effect of its own output, and to modify its behaviour according to its internal model of the listener.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 213850 - SCALE.

7. REFERENCES

[1] Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.
[2] Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11), 1039-1064.
[3] Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (Eds.). (2001). *Improvements in Speech Synthesis*. Chichester, UK: Wiley & Sons.
[4] Lombard, E. (1911). Le sign de l'élévation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37, 101-119.
[5] Junqua, J.-C. (1996). The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20, 13-22.
[6] Moore, R. K. (2003). Spoken language output: realising the vision, *EUROSPEECH*. Geneva.
[7] Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403-439): Kluwer Academic Publishers.
[8] Moore, R. K. (2006). Spoken Language Processing for Artificial Cognitive Systems, Keynote talk at the session on Cognitive Systems, Interaction and Robotics, *EU IST Conference*. Helsinki.
[9] Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. *Speech Communication*, 49, 418-435.
[10] Moore, R. K. (2007). PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Trans. Computers*, 56(9), 1176-1188.
[11] Powers, W. T. (1973). *Behaviour: The Control of Perception*. NY: Aldine: Hawthorne.
[12] Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14, 41-104.
[13] Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, 77(2), 97-132.
[14] Tang, Y., Cooke, M. (2010). Energy reallocation strategies for speech enhancement in known noise conditions, *INTERSPEECH* (pp. 1636-1639). Makuhari, Japan.
[15] Picart, B., Drugman, T., Dutoit, T. (2010). Analysis and synthesis of hypo and hyperarticulated speech, *7th ISCA Speech Synthesis Workshop*. Kyoto, Japan.
[16] Hazan, V., & Simpson, A. (1998). The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24(3), 211-226.
[17] www.cstr.ed.ac.uk/projects/festival/
[18] Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., 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(6), 1208-1230.
[19] www.synsig.org/index.php/Blizzard_Challenge_2010
[20] ANSI (1997). ANSI S3.5-1997, *American National Standard Methods for Calculation of the Speech Intelligibility Index*. American National Standards Institute, New York.