Altering speech synthesis prosody through real time natural gestural control

David Abelman[1]*, Robert A.J. Clark*[2]

[1]School of Informatics, University of Edinburgh, Edinburgh, UK
[2]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
robert@cstr.ed.ac.uk

Abstract

This paper investigates the usage of natural gestural controls to alter synthesised speech prosody in real time (for example, recognising a one-handed beat as a cue to emphasise a certain word in a synthesised sentence). A user's gestures are recognised using a Microsoft Kinect® sensor, and synthesised speech prosody is altered through a series of hand-crafted rules running through a modified HTS engine (pHTS, as described in [1]). Two sets of preliminary experiments are carried out. Firstly, it is shown that users can control the device to a moderate level of accuracy, though this is projected to improve further as the system is refined. Secondly, it is shown that the prosody of the altered output is significantly preferred to that of the baseline pHTS synthesis. Future work is recommended to focus on learning gestural and prosodic rules from data, and in using an updated version of the underlying pHTS engine.

The reader is encouraged to watch a short video demonstration of the work at ***http://tinyurl.com/gesture-prosody***.

Index Terms: speech prosody, pHTS, real time control, gesture

1. Introduction

Despite the significant advances made in speech technology in recent years, producing speech that is both **expressive** and **reactive** to a user's input or local environment is a significant challenge facing speech synthesis today, and one on which there has been limited research to date [1] [2].

This work aims to construct a system to alter speech prosody in a limited number of ways, based on a user's real time gestural input. Future extended systems of this type would have various potential applications. A primary use may be in text-to-speech communication aids of those with vocal disorders. More natural expression and prosody may be 'conducted' by the user in real-time, either through a set of standard natural gestures, or through a set of custom-designed gestures for those with physical disabilities (for example, eyebrow or finger movements).

Additionally, technology developed as part of this system may be incorporated within potential 'sign-language synthesis' systems of the future [3] [4]. In addition to synthesising words based on sign-language hand movements, the manner in which the gestures are performed may indicate to the system a certain expressive or emphatic style in which to synthesise the speech.

Finally, other potential applications may exist within the entertainment industry. For example, the technology may be adapted for use within synthesised singing voices, or perhaps within future 'instrument-voice hybrids' that people may wish to control through body gestures. Ultimately, any situation in which it would be useful to improve expressiveness of a voice-like synthesis in real-time would benefit from the research that this work undertakes.

2. Background

This project is built upon a modified HTS engine 'pHTS' (*performative* HTS) [1]. This technology allows HTS synthesis to be reactive to its environment - whether adapting to surrounding conditions, or being controlled expressively by a user (as is the case here). In order to make the system reactive, the phonetic context required in calculating the synthesis parameters is reduced from that of the whole sentence to a much smaller window. This change requires two main modifications. Firstly, the context used in training the model is reduced to just the current and surrounding phonemes, and the current and previous syllable. Secondly, during synthesis, the generation of parameters occurs on a sliding window of two labels.

A variety of potential applications for pHTS have been outlined and developed by the group. These include *HandSketch*, a pen-based musical instrument prototype [5], speech synthesis based on face-tracking [6], and accent interpolation through an interactive map application [7]. Meanwhile, non-pHTS examples of hand-controlled prosodic modification include [8].

Finally, [2] incorporates skeleton tracking (using Microsoft Kinect) into the pHTS system to create a reactive speech synthesiser, in which pitch and duration are controlled by the vertical position of both hands. It is found that meaningful expressiveness is difficult to simulate when pitch and duration modulations are controlled in this particular way. It is this work that this project intends to build on, incorporating gesture recognition and more constrained prosodic modification rules.

3. Design

3.1. Preliminary decisions

A number of preliminary decisions were made to constrain the scope of the project. Future iterations of the work should revisit these in order to extend the system's abilities. Decisions included:

- Limiting the system's prosodic vocabulary to contrastive emphasis, general emphasis, yes/no questions and wh-questions only
- Realisation of prosodic shifts through manual parameter shifts on a single speech database, as opposed to switching between multiple recorded databases for different effects without explicitly shifting speech parameters
- Alteration of pitch and duration only, as the two primary parametric drivers of prosody (i.e. volume, spectral energy, pause models etc. are left untouched)

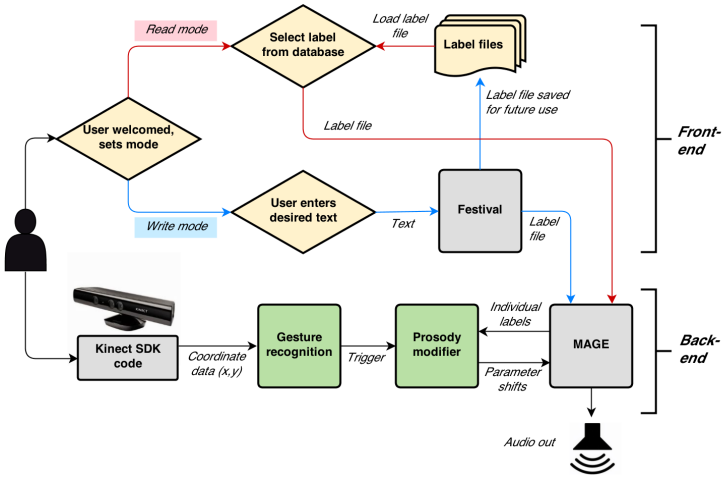

Figure 1: *System setup.*

- Magnitudes of pitch/duration shifts are manually encoded, as opposed to being learnt through data
- Gesture recognition implemented in a rule-based fashion, as opposed to being learnt through data
- Limiting the system's gesture recognition to one handed beats (contrastive and general emphasis, different hand for each), head tilts (yes/no questions and wh-questions, different side for each)
- Constricting emphasis to content words (not function words) and to stressed syllables within these words only. Future iterations may use a more flexible or intelligent natural language model to bias the prosodic effects towards the most likely syllables given the semantic meaning of the text, and the user's timing

A number of small-scale tests were carried out in order to optimise values for parameter shift magnitudes, and gesture tracking rules. Details may be found in [9]. The final set of rules implemented in terms of skeletal coordinates and pHTS parameter shifts are laid out in Section 3.3.4 of [9], with an illustrative example (contrastive emphasis) shown in this paper in Section 3.3.

3.2. System setup

A schematic for the system set up is laid out in Figure 1. The system has two modes, **read** and **write**. Both modes create a temporary label file which the backend loads into MAGE. **Write** mode passes the user's textual input (currently entered using a keyboard) to **Festival** Speech Synthesis System v2.1[1] in order to create the label file, whereas **read** mode allows the user to select a label file previously created by this process. The format of the label file required by the system is that outlined within [10].

The backend is built in C++ on top of MAGE and the work already carried out as part of [2]. OpenFrameworks[2] is used as the graphical and audio framework. The main application loop repeatedly calls an 'update' function (15 times per second), in which pre-existing **Kinect SDK code** tracks *x* and *y* coordinate data for the user's skeleton.

[1]http://www.cstr.ed.ac.uk/projects/festival/download.html
[2]http://www.openframeworks.cc/

The skeletal coordinate data are used within the **gesture recognition** stage. A simple set of if-else rules act as triggers for initiating prosodic effects, as described in [9]. Once some gesture has been recognised, the **prosody modifier** must allocate the prosodic effect to a specific syllabic unit (or set of syllabic units). Prosodic adjustments are carried out using a set of if-else rules, as also described in [9].

The parameter shifts are sent to the **MAGE** engine, and are used to shift parameter trajectories as appropriate via the pHTS engine. This functionality pre-exists within the MAGE platform code.

In addition to the resulting audio output, the application provides a visual representation of the user's skeleton as part of the application interface. Flashing text and graphical meters indicate to the user when a gesture has been recognised, and how the pitch and duration of the synthesis is being shifted.

3.3. Contrastive emphasis example

This section briefly outlines the set of rules implemented to enable the system to add *contrastive emphasis* to a synthesised sentence. Similar sets of rules exist for *general emphasis*, *yes/no questions* and *wh-questions*, and can be found in [9].

Gestural rules: As the Kinect recognises the left hand moving above the left hip, a 'contrastive window' of ∼0.5 seconds (8 frames) is triggered. (Note that due to latency issues it has not been possible to fine control the placement of this window for any hand movements above the hip, though this would be desirable in an improved system).

Prosodic rules: A pitch accent is applied if a content word's stressed syllable falls within this 0.5 second window. The pitch accent consists of a raised pitch (28%) and a reduction in speed (-10%). Following this, the remainder of the sentence is lowered in pitch (-14%) and increased in speed (14%). The final syllable of the sentence is raised in pitch (10%) to counter the default falling accent provided by pHTS.

4. Experiment 1 - Generation test

Two experiments have been carried out, the first being a generation test to investigate the accuracy to which users can control the system.

4.1. Experimental setup

In total, 12 native English speakers were tested using the system. Each user was asked to add some form of emphasis to 31 different synthesised sentences through gestural control, with each sentence being repeated eight times consecutively. Each sentence contains one or two gestures, leading to a total of 264 gestures that have been performed and tracked for each user, taking around 50 minutes in total.

A script was placed within the user's view. For each sentence the user was told which word to emphasise, and with what action (normally contrastive emphasis, as this is the most obvious to the ear). The author was able to track each attempt as being correct, early / late (missing the intended syllable but not emphasising an unintended syllable) or very early / very late (emphasising an unintended syllable). Immediately after each attempt and prior to the next, the subject was told by the author if they had gestured correctly, early or late (although it was often already clear to the user without prompting). This feedback is justified, as we would expect a fully-developed system to pro-

vide the user with some kind of explicit feedback on where their attempted emphasis fell.

The sentences are split into five primary sections. These sections are presented to each subject in the same order (avoiding learning bias over the session), but sentences within each section are presented in different orders according to various Latin Squares. Full sentence lists may be found in the original work's appendix [9].

4.2. Experimental results

A selection of findings are presented here. Two-tailed binomial tests are used to mark 95% confidence intervals in tables. The mean of the binomial distribution is set to the proportion of correct emphases out of all attempts. Chi-squared tests are used to calculate p-values for significance when confidence intervals overlap.

Spread of false positive and negative gesture timings: Considering just the *first attempt* across all sentence types in the 50 minute experiment, correct emphasis is applied 50% of the time. The user emphasises *no* word 30% of the time (gesturing only slightly too late or early), and the *wrong* word 20% of the time. These results are shown in Table 1. Note that when *all eight attempts* are considered, the application of correct emphasis improves from 50% to 65% (as users improve with practice on each sentence).

Table 1: *Spread of emphasis gesture timings (1st attempt only)*

Emphasis	% of time
Very early (wrong emphasis)	$7 \pm 3\%$
Early (no emphasis)	$4 \pm 2\%$
Correct emphasis	$50 \pm 6\%$
Late (no emphasis)	$26 \pm 5\%$
Very late (wrong emphasis)	$12 \pm 4\%$

Improvement over session: Users were requested to add emphasis onto individual words within a sentence at the start of the experiment, and onto individual words in sentences of similar rhythm after 25 and 45 minutes of practice. Results show that users do improve between 0 and 25 minutes ($p<0.01$), though not significantly between 25 and 45 minutes, suggesting that accuracy levels plateau with reasonably little experience. These results are shown in Table 2.

Table 2: *Accuracy improvement over the session*

Time since experiment start	Avg. accuracy (8 attempts)
0 minutes	$52 \pm 7\%$
25 minutes	$64 \pm 7\%$
45 minutes	$67 \pm 7\%$

Other results: Other results found include the following:

- A user emphasising *two words per sentence* will on average obtain a lower accuracy rate for the second word, in comparison to emphasising that same word *alone* in a sentence, if the words are in close enough proximity.
- Users have a significantly lower accuracy rate emphasising *words at the start and end of sentences*. Emphasising a word at the beginning of a sentence may be expected to be more challenging, as the user can be caught off-guard. There is less clear reason for words at the ends of sentences to be harder to emphasise - this may be due to a quirk of the sentences chosen within this experiment (for example, the 'rhythm' with which the synthesiser recites them).
- A user *speaking the text out loud* at the same time as gesticulating to control the synthesiser does not find his/her accuracy altered significantly, on average. It had been hypothesised that gestures may be performed more naturally at the correct moments if the user was speaking out loud whilst gesticulating. However, given this result, this wouldn't be a technique that is recommended to users in any future system.
- The *naturalness* (or unnaturalness) of the word to be emphasised does not affect a user's accuracy rate significantly.

5. Experiment 2 - Listening test

A listening test has been carried out to investigate the extent (if any) by which the output is perceived to be more natural, or have a different meaning, relative to baseline pHTS. Null hypotheses assumes a listener chooses an option from the forced choice test with equal probability - i.e. the options are equivalent. Two-tailed binomial tests are used to calculate p-values, and 95% confidence intervals are shown in tables.

5.1. Experimental setup

In total, 33 subjects were recruited for a listening test, lasting 20-30 minutes depending on the subject, under controlled conditions. All subjects identified themselves as being native English speakers, and received £6 in compensation. Each user was presented with 92 sentences split across 7 sections, and asked to select one of two options in a forced-choice test. This choice involved selecting either a preferred audio clip or a preferred textual option, depending on the question. Similarly to the generation test, sections were presented in a consistent order, but questions and options within each section were appropriately randomised. Once again, a full list of test sentences may be found in the original work's appendix [9].

5.2. Experimental results

A selection of findings regarding **contrastive emphasis** are presented here.

Perceived naturalness of contrastive emphasis: A question narrated by the author and pairs of synthesised responses were played to listeners. The sentences were of a defined form - see [9] for details. One of the responses would be a neutral pHTS synthesis, the other would have a word emphasised using the reactive synthesis system. This emphasis may be appropriate (correct word emphasised) or inappropriate (incorrect word emphasised). The participant must choose which of the two responses seems more natural. To illustrate, the user was hypothesised to prefer the 'appropriately' emphasised (first) response in this case:

'Did Jess have trout for her breakfast yesterday?'

1. *'No, Jess had **SALMON** for her breakfast yesterday.'*
2. *'No, Jess had salmon for her breakfast yesterday.'*

whereas the user was hypothesised to prefer the neutral pHTS baseline (second) response where an 'inappropriately' emphasised response was presented:

'Did Jess have trout for her breakfast yesterday?'

1. *'No, Jess had salmon for her **BREAKFAST** yesterday.'*
2. *'No, Jess had salmon for her breakfast yesterday.'*

Indeed, it has been found that listeners significantly prefer contrastive emphasis over neutral prosody when the emphasis is delivered on the appropriate word, and significantly prefer the neutral prosody over contrastive emphasis when the emphasis is delivered on an inappropriate word ($p<0.01$). Both of these results are as hypothesised, and are shown in Table 3.

Table 3: *Contrastive emphasis - naturalness*

Listener preference:	Emphasised	Neutral
Appropriately emph'd synthesis	**86 ± 4%**	14 ± 4%
Inappropriately emph'd synthesis	22 ± 7%	**78 ± 7%**

Effect on semantic interpretation of sentence by contrastive emphasis: Two textual options were presented to participants in writing, along with a single synthesised audio response. For example, a pair of textual options used was:

1. *The black dog was lying on the mat*
2. *The white mouse was lying on the mat*

with the single synthesised audio response being:

*'No, the **WHITE** dog was lying on the mat.'*

Since the emphasis is on *'WHITE'*, we would expect the listener to select the first of the two textual statements as the more appropriate, given the response. (If the emphasis had been on *'DOG'* we would have expected the user to choose the second textual option). Indeed it was found that these expected choices were made the majority of the time: the position of the synthesised emphasis significantly changes the user's semantic interpretation of the response ($p<0.01$). Results are shown in Table 4.

Table 4: *Contrastive emphasis - semantics*

Emphasis perceived to be on:	1st word	2nd word
When neutral synthesis	32 ± 7%	**67 ± 7%**
When 1st word emph'd in synthesis	**94 ± 6%**	5 ± 6%
When 2nd word emph'd in synthesis	8 ± 5%	**92 ± 5%**

Other results: Other results found include the following:

- As described above, within a defined sentence structure, appropriate contrastive emphasis is considered more natural than neutral synthesis. However, a separate section of the listening test showed that appropriate contrastive emphasis placed on the *final* syllable of a sentence is *not* perceived to be more natural by listeners. The hypothesised reason is that a final syllable emphasised contrastively needs to rise *and* fall in pitch within the same syllable. However, contrastive emphasis built into the current system only raises the emphasised syllable, resulting in an unnatural effect. This issue should be addressed in any future iterations of the work.
- Although contrastive emphasis was set as described in Section 3.3 using a 28% pitch rise on the emphasised syllable for most of the listening test, alternative values for the magnitude of the pitch rise were evaluated against one another within a short section of the test. A rise of roughly 20% was found to be optimal according to listeners' choices. Future iterations of the system may tweak parameters through tests such as these to optimise perceived naturalness.
- The experiment also evaluated listeners' perceptions of interrogative prosody, in addition to contrastive emphasis as outlined here. The reader may consult the original work for details [9].

6. Discussion

This work presented does suggest that it is possible to improve prosody of speech synthesis in real time through gestural controls. Users can control the emphasis with some accuracy, and listeners overwhelmingly prefer correctly emphasised sentences over baseline pHTS. It should be noted however that no analysis of the benefit of a correct emphasis versus the cost of an incorrect emphasis (from the listener's point of view) has been carried out within this work.

Future work should enable the user to control the system with a much superior accuracy rate to that obtained here. The largest obstacle to accurate control within this iteration of the work was relatively poor latency, caused by audio buffering. Future iterations of the work will use MAGE 2.0 [11] rather than MAGE 1.0, which the current system is based on. This will improve audio buffering times, meaning that the system has the potential to react to more granular gesture timings. For example, rather than triggering a window for potential emphasis as a user raises their wrist above their hip, the system may apply emphasis within an instant of recognising that the user's hand has reached the apex of an emphatic 'beat' trajectory. This should improve accuracy rates significantly. Additionally this would allow prosody to be affected *before* the user reaches this apex if required, for example by decreasing the speed of a syllable immediately preceding a contrastively emphasised syllable.

Future iterations of the system should implement machine learning based techniques for gestural recognition, as opposed to the current rule-based setup. This will result in a more flexible system going forward, in which new gestures can be recorded more easily, added or modified by a user, and customised to those with accessibility requirements. Similarly, learning prosodic parameter shifts through data (as opposed to the hard-coded rules currently used) will allow the system to scale more easily to a larger repertoire of available prosodic effects. A final modification that may improve the system's accuracy would be to incorporate a discourse model to aid prediction of where the user intends to apply emphases and similar prosodic effects. Even if the user's timing is slightly out, the system may then intelligently factor in prior probabilities of words most likely to be emphasised given the discourse context.

In summary, the work presented is in its early stages, but improvements such as these will lead to a novel and natural method of altering speech synthesis prosody in real time.

7. References

[1] Maria Astrinaki, Nicolas D'alessandro, Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Reactive and continuous control of HMM-based speech synthesis. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 252–257. IEEE, 2012.

[2] Robert AJ Clark, Magdalena Anna Konkiewicz, Maria Astrinaki, and Junichi Yamagishi. Reactive control of expressive speech synthesis using Kinect skeleton tracking. *Information Processing Society of Japan*, 112(369):175–178, 2012.

[3] S Sidney Fels and Geoffrey E Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *Neural Networks, IEEE Transactions on*, 4(1):2–8, 1993.

[4] Robert Akl. *Evaluating appropriateness of EMG and flex sensors for classifying hand gestures*. PhD thesis, University of North Texas, 2012.

[5] Maria Astrinaki, Nicolas dAlessandro, and Thierry Dutoit. MAGE - A platform for tangible speech synthesis. In *Proceedings of the 12th Conference on New Interfaces for Musical Expression (NIME'12)*, 2012.

[6] Maria Astrinaki, Nicolas D'alessandro, and Thierry Dutoit. MageFaceOSC: Performative speech synthesis based on realtime face tracking. *QPSR of the numediart research program*, 5(1):15–16, 2012.

[7] Maria Astrinaki, Junichi Yamagishi, Simon King, Nicolas dAlessandro, and Thierry Dutoit. Reactive accent interpolation through an interactive map application. *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, 2013.

[8] Christophe dAlessandro, Albert Rilliard, and Sylvain Le Beux. Chironomic stylization of intonation. *The Journal of the Acoustical Society of America*, 129(3):1594–1604, 2011.

[9] David Abelman. Altering speech synthesis prosody through real time natural gestural control. *Edinburgh University MSc Thesis (http://hdl.handle.net/1842/8373)*, 2013.

[10] Heiga Zen. An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, 2006.

[11] Maria Astrinaki, Nicolas D'alessandro, Loic Reboursière, Alexis Moinet, and Thierry Dutoit. MAGE 2.0: new features and its application in the development of a talking guitar. In *Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, 2013.