Reactive accent interpolation through an interactive map application

Maria Astrinaki[1], *Junichi Yamagishi*[2,3], *Simon King*[2], *Nicolas d´Alessandro*[1], *Thierry Dutoit*[1]

[1]Circuit Theory and Signal Processing Lab, University of Mons, Belgium
[2]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[3]National Institute of Informatics, Tokyo, Japan

maria.astrinaki@umons.ac.be, jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk
nicolas.dalessandro@umons.ac.be, thierry.dutoit@umons.ac.be

Abstract

In this paper, we present our first prototype system for interactive accent control using HMM-based speech synthesis. In this application, voices in various English accents including American, Canadian and British, are controlled and interpolated by the user using gestures acquired via an interactive geographical map in real time. Users can choose the gender of the voices to manipulate and the accent interpolation strategy: either interpolation of all available speakers within an area or the N-nearest speakers around a point.
Index Terms: speech synthesis, reactive, dialect, interpolation

1. Introduction

The flexibility of statistical parametric speech synthesis based on hidden Markov models (HMMs) is well-established, e.g., [3, 1]. All aspects of speech, comprising the spectral envelope, excitation parameters, and segment duration are modeled and generated simultaneously in one framework. A significant advantage of this model-based parametric approach is the ability to manipulate the model parameters in principled ways – such as adaptation to new data or interpolation between models – and thus manipulate the characteristics of the generated speech. These techniques have already been applied to generating or morphing between different speakers, different types of emotional speech, and different speaking styles [3].

In this paper, we will use the word *accent* to mean the pronunciation of individual vowels and consonants (such as the height of a particular vowel). We will assume there is, at least to a first approximation, a continuum of accents: that is, accent changes gradually in accordance with other speaker factors such as geographical location. Therefore, since the acoustic realisation of phonemes gradually changes, accent can be manipulated through the interpolation of models mentioned above. In other words, given two voices with differing accents, it is reasonable to create an accent "halfway" between them.

Varying the *dialect* of a text-to-speech (TTS) synthesis is more complex because this can involve changes beyond accent, into the phonology, vocabulary and grammar of the language; it is no longer reasonable to assume there is a continuum and one must deal with discontinuities, such as abrupt changes in the phonology (e.g., the number of phonemes in the language). An example of this can be found in [4].

In this paper, we present our first prototype system for interactive accent control using HMM-based speech synthesis. Various English accents including American, Canadian and British, can be controlled and interpolated by using the user's gestures acquired over an interactive geographical map in real time. Users can choose the gender of the voice and the interpolation strategy, which is either an interpolation of all available speakers within an area specified by the user or the N-nearest speakers around a point specified by the user. In Section 2, we describe the platform used for reactive HMM-based speech synthesis. Then, we present the interactive map application, an interface for reactively controlling accent interpolation through gesture, in Section 3. For simplicity of implementation, in the present work at least, we do not explicitly consider discontinuities in accent.

2. Reactive HMM-based speech synthesis

In the application, various English accents need to be controlled and interpolated in a real time fashion. Therefore we used MAGE [2], which is a real time architecture for reactive HMM-based speech and singing synthesis that allows user control over the speech waveform at various levels. Figure 1 illustrates the threaded architecture of MAGE which comprises three main threads: the label thread, that is responsible for the contextual control, the parameter generation thread, that generates the spectral and excitation parameters for only a single phoneme and finally the audio generation thread, that generates the speech samples corresponding to the inputed phoneme. All threads can be reactively controlled by the user.

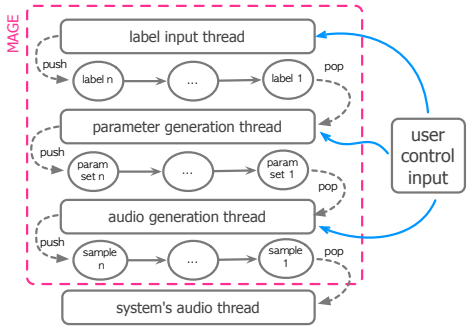

Figure 1: Multithread architecture of MAGE: all threads can be reactively controlled by the user.

3. Reactive accent interpolation through an interactive map application

For the reactive accent interpolation, it is important to separate out speaker characteristics and accent somehow so listeners can pay attention only to accent transitions. One choice might be to use a single speaker who can produce all accents

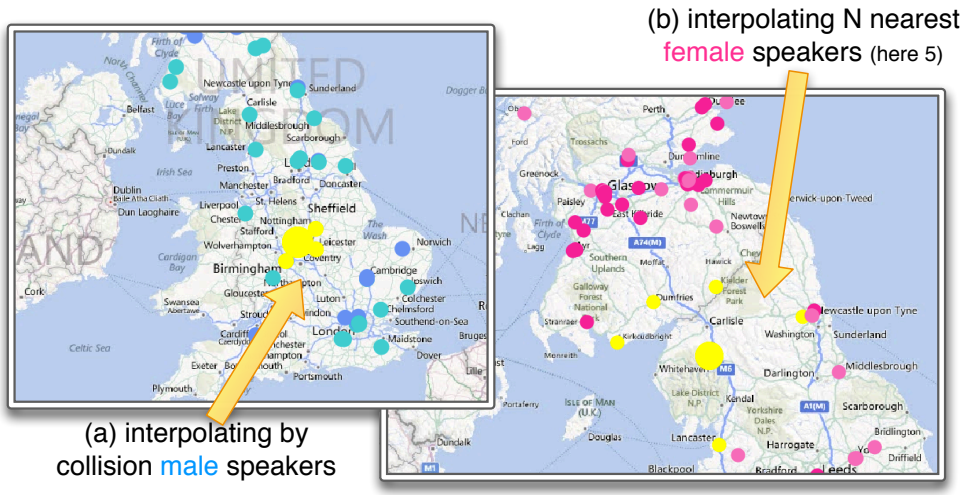

Figure 2: Examples of interactive accent interpolation during (a) "collision" and (b) "continuous" modes.

correctly. However, this is impractical: it would he hard or impossible to find such a speaker. Instead, we decided to use multiple speakers who have similar accents, by interpolating their acoustic models. In the application, the listeners only hear such an interpolated voice (i.e., never that of a single speaker). As they interact with the map, the perceived accent will gradually change depending on the particular speakers being interpolated. All speakers are chosen from the CSTR voice banking corpus, which contains a total of around 400 speakers with many and varied UK accents.

3.1. Geographical map

The application consists of the world map, on which every single speaker is represented as a circle. The "active" region controlled by the user is represented as a yellow circle around the cursor. There are zoom in and zoom out functions, as well as navigational functions, in order to navigate through the map. These controls are available with mouse scroll up or down and with clicking and dragging respectively. A touch screen can be employed to provide a more intuitive experience for users. In addition, there is a switch to change the gender of the speakers.

There are two interpolation strategies: "collision" and "continuous". During the collision interpolation mode, when the yellow active region overlaps with one or more speaker circles then the selected speakers will be interpolated and used for speech synthesis. When the cursor does not overlap with any speaker, a default voice is used. During the continuous mode, each time the cursor moves, the distance between the cursor and all the available speakers is computed and the N-nearest neighbouring speakers are selected and interpolated for speech synthesis. Figure 2 shows examples of the two interactive accent interpolation modes in operation.

3.2. Speech synthesis

The waveform generation part of the application is implemented by the MAGE reactive speech synthesizer. When the list of targeted speakers for interpolation is received, the corresponding voice models are loaded into the engine. Then, the interpolation weights are computed; we use uniform weights of $w = 1/N$ for each voice model. After interpolating the loaded models using the computed weights, the speech parameter trajectories are generated and the speech output is synthesized.

3.3. Potential applications

The interactive accent map application could have several applications, targeting the creation and use of unique personalized voices. In the field of new interfaces for musical expression and performing arts it could be possible to create voices with specific accents suitable for performance. In gaming applications users could customize and refine the accent of their avatar. In GPS applications the accent of the voice used could be adapted depending on its position. Considering movie dubbing applications or assistive applications for speech impaired people, voices could be adapted and personalized to individual people. A specific example of this would be to allow people who cannot speak for themselves to create their own unique voice (or voices!). There could be interesting applications in speech pedagogy and therapy by creating adaptive references for certain dialects. Linguists might investigate how accent changes with geography, and investigate the existence of isoglosses.

4. Conclusions

It is not straightforward to formally evaluate the proposed interactive accent control, but we can certainly demonstrate that the manipulation of accent in real time is feasible. However, a few aspects of the system would benefit from improvement. Currently the manipulation of accent is phoneme-based and no restriction as to how and when modifications could be made were imposed. Setting more restricted patterns to convey the final dialect, such as the ones used in natural human speech, could probably lead to more linguistically distinguishable results.

5. References

[1] H. Zen, K. Tokuda and A. W. Black, Statistical Parametric Speech Synthesis, Speech Communication, 1039–1064, 51, 2009.

[2] [Online] Mage Platform for Performative Speech Synthesis, M. Astrinaki and A. Moinet and G. Wilfart and N. d´Alessandro and T. Dutoit, http://mage.numediart.org/.

[3] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, Speech Synthesis with Various Emotional Expressions and Speaking Styles by Style Interpolation and Morphing, IEICE Transactions, E88-D, 11, 2484–2491, 2005.

[4] M. Pucher, D. Schabus, J.Yamagishi, F. Neubarth, and V. Strom, Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis, Speech Communication, 52, 2, 164–179, 2010