Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 176–193

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis

Markus Toman [a,*], Michael Pucher [a], Sylvia Moosmüller [b], Dietmar Schabus [a]

[a] Telecommunications Research Center Vienna (FTW), Donau-City-Str 1, 3rd floor, 1220 Vienna, Austria
[b] Austrian Academy of Sciences – Acoustics Research Institute (ARI), Wohllebengasse 12-14, 1st Floor, 1040 Vienna, Austria

Received 10 December 2014; received in revised form 18 May 2015; accepted 4 June 2015
Available online 12 June 2015

Abstract

This paper presents an unsupervised method that allows for gradual interpolation between language varieties in statistical parametric speech synthesis using Hidden Semi-Markov Models (HSMMs). We apply dynamic time warping using Kullback–Leibler divergence on two sequences of HSMM states to find adequate interpolation partners. The method operates on state sequences with explicit durations and also on expanded state sequences where each state corresponds to one feature frame. In an intelligibility and dialect rating subjective evaluation of synthesized test sentences, we show that our method can generate intermediate varieties for three Austrian dialects (Viennese, Innervillgraten, Bad Goisern). We also provide an extensive phonetic analysis of the interpolated samples. The analysis includes input-switch rules, which cover historically different phonological developments of the dialects versus the standard language; and phonological processes, which are phonetically motivated, gradual, and common to all varieties. We present an extended method which linearly interpolates phonological processes but uses a step function for input-switch rules. Our evaluation shows that the integration of this kind of phonological knowledge improves dialect authenticity judgment of the synthesized speech, as performed by dialect speakers. Since gradual transitions between varieties are an existing phenomenon, we can use our methods to adapt speech output systems accordingly.

Keywords: HMM-based speech synthesis; Interpolation; Austrian German; Innervillgraten dialect; Bad Goisern dialect; Viennese dialect

1. Introduction

The flexibility of Hidden Semi-Markov Model (HSMM) based speech synthesis allows for different strategies to manipulate the trained models, such as adaptation and interpolation. In this paper we develop, analyze, and evaluate unsupervised interpolation methods that can be used to generate intermediate stages of two language varieties. "Variety" is a cover term void of any positive or negative evaluative connotations. It comprises dialects, sociolects, and standard languages. In this contribution, we apply this method to perform an interpolation between Regional Standard Austrian German (RSAG) and three dialects/sociolects. The difficulty of dialect interpolation lies in lexical, phonological, and phonetic differences between the varieties (Russell et al., 2013). In this contribution we focus on interpolation of phonetic differences.

In recent years there have been several research efforts in the context of language varieties for speech synthesis, reviewed in Russell et al. (2013). Following Russell et al. (2013) we can distinguish between fully-resourced and under-resourced modeling as well as different applications like variety interpolation.

In fully-resourced modeling, Richmond et al. (2010) described how to generate pronunciation dictionaries based

* Corresponding author.
E-mail address: toman@ftw.at (M. Toman).

http://dx.doi.org/10.1016/j.specom.2015.06.005
0167-6393/

on morphological derivations of known words. They reported that in preliminary experiments for 75% of tested words, their method produced the correct, fully-specified transcription. This can be used as an extension to existing grapheme-to-phoneme rules to obtain contextual information on out-of-vocabulary words and could be beneficial for building an actual dialect synthesis system that includes interpolation.

Nguyen et al. (2013) described the development of an HMM-based synthesizer for the modern Hanoi dialect of Northern Vietnamese, describing special challenges they encountered, comparable to our process of acquiring our dialect corpus.

In Toman et al. (2013b) we evaluated different acoustic modeling methods for dialect synthesis. The interpolation technique presented in the present work is compatible to all acoustic modeling methods as long as they produce a HSMM state sequence for a given set of labels.

For developing synthesizers for under-resourced languages, different methods have been developed to aid the process of data acquisition and annotation.

Goel et al. (2010) evaluated the combination of different lexicon learning techniques with a smaller lexicon available for bootstrapping. In their experiments, their method could increase the Word Recognition Accuracy from 41.38% for a small bootstrap lexicon to 43.25%, compared to 44.35% when using the full training dictionary.

Watts et al. (2013) developed methods and tools for (semi-)automatic data selection and front-end construction for different languages, varieties and speaking styles e.g. from audio books. Results from Watts et al. (2013) are published by Stan et al. (2013) who applied these tools on "found speech" to create a standardized multilingual corpus. For our work on dialectal synthesis, such methods are useful for easy acquisition and annotation of dialect data, which is currently a time-consuming process.

Loots and Niesler (2011) developed a phoneme-to-phoneme conversion technique that uses decision trees to automatically convert pronunciations between American, British and South African English accents.[1] This method could be used to automatically generate the phonetic transcription for less-resourced dialects from a fully-resourced variety, as a transcription of the dialect utterance is required for our interpolation technique presented here.

Voice model interpolation was first applied in HSMM-based synthesis for speaker interpolation (Yoshimura et al., 2000) and emotional speech synthesis (Tachibana et al., 2005). Picart et al. (2011) used model interpolation to create speech with different levels of articulation. Lecumberri et al. (2014) considered the possibility of using extrapolation techniques to emphasize foreign accent as an application for foreign language learning. The methods presented here could also be used to produce an extrapolated dialect, but this is not investigated in the current paper.

In language variety interpolation, Astrinaki et al. (2013) have shown how to interpolate between clusters of accented English speech within a reactive HMM-based synthesis system. In this method, phonetic differences between the accent representations were not considered (i.e. the same set of phone symbols and utterance transcriptions was used for all accents).

In Pucher et al. (2010), we have shown how to interpolate between phonetically different dialects in a supervised way. In this method, we used a manually defined phone-mapping between Standard Austrian German and the Viennese dialect. Evaluation tests showed that listeners actually perceive the intermediate varieties created by interpolation as such.

In this contribution we extend the method from Pucher et al. (2010) to work in an unsupervised way, such that no manually defined mapping is necessary, therefore allowing the fully automatic interpolation. Also, interpolation is performed between RSAG and three dialects/sociolects. This unsupervised method is based on Dynamic Time Warping (DTW) (Rabiner et al., 1978) on HSMM state level and is subsequently described in Section 3. Compared to Pucher et al. (2010), this method introduces one-to-many mappings between states, requiring a more sophisticated duration modeling procedure, which will be described in Section 4.

To introduce the integration of phonological knowledge in the interpolation technique, we describe the following alternations, which characterize the RSAG – dialect interaction[2]:

1. **Phonological process**: Socio-phonological studies on Austrian varieties demonstrate that certain alternations between two varieties, usually a standard variety and a dialect, are phonetically well motivated and thus can be described as phonological processes, e.g., spirantization of intervocalic lenis stops (Moosmüller, 1991) like
   - [ɑːb̥ɐ] to [ɑːbɐ] to [ɑːβɐ]
     aber (engl. "but") or
   - [laɛd̥ɐ] to [laɛdɐ] to [laɛäɐ]
     leider (engl. "unfortunately").

   Interpolation can be used to model these gradual transitions.
2. **Input-switch rules**: Other alternations lack such phonetic motivations because of a different historical development. These alternations are therefore described as input-switch rules, e.g.

[1] The term "accent" is often used for regional differences of English. We avoid the term "accent" in this contribution as it refers to more than one linguistic phenomenon and we specifically treat dialects here.

[2] // denotes the phonological representation, [] the phonetic realization.