An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements

Sacha Krstulović, Anna Hunecke and Marc Schröder

DFKI GmbH, Saarbrücken, Germany

{sacha.krstulovic,anna.hunecke,marc.schroeder}@dfki.de

Abstract

The paper assesses the capability of an HMM-based TTS system to produce German speech. The results are discussed in qualitative terms, and compared over three different choices of context features. In addition, the system is adapted to a small set of football announcements, in an exploratory attempt to synthesise expressive speech. We conclude that the HMMs are able to produce highly intelligible neutral German speech, with a stable quality, and that the expressivity is partially captured in spite of the small size of the football dataset.

Index Terms: HMM-based speech synthesis, synthesis of German speech, expressive speech synthesis.

1. Introduction

In state-of-the-art unit-selection speech synthesis, the expressiveness of the synthetic speech is rigidly linked to the contents of the underlying database. To reach a better expressivity, either the database should be made larger, with a related increase in the costs of database querying and storage, or a method should be found to parameterise the expressiveness of any given inventory of speech units, in order to interpolate or extrapolate the possibly unseen expressive units. Our research on this topic has led us to question which model would abstract the speech units with enough flexibility and detail to support an adequate control of the prosody and voice quality in relation to expression variations, while keeping a good level of perceptual quality.

In this perspective, Hidden Markov Models (HMMs) have proven to be an efficient parametric model of the speech acoustics in the framework of speech synthesis [1]. Furthermore, HMMs and Gaussian modelling have proven that they can support class-dependent transformations of the speech acoustics, while reaching a competitive perceptual quality. This is illustrated, e.g., by the fact that methods based on Gaussian Mixture Models have gradually emerged as the state-of-the-art in the domain of cross-speaker voice transformation [2, 3]. Along this line of evolution, the adaptation of an HMM-based synthesis system to an arbitrary speaker's voice [4] has been implemented through linear adaptation methods inherited from speech recognition. These methods transform an average model, trained over a large speech sample comprising several speakers and a large sampling of the linguistic space, into a speaker-specific model, via a set of linear transforms that are learnt over a limited sample of speaker-specific data. Moreover, a set of class-dependent HMMs can be used as a model space across which some acoustic transformations can take place. Examples of this logic include model interpolation, which has been applied to obtain gradual transformations between the voices of different speakers [5], and eigenvoices [6], which aim at reducing the cross-model variations to a limited number of control parameters.

Our interest is to apply similar modelling and transformation techniques across classes of speech expressivity rather than classes of speaker identity, the target application being the synthesis of emotional speech in German. In this domain, recent results have indicated that the HMM-based systems are able to produce speech in different speaking styles for the Japanese language [7], and that an explicit parameterisation of the speaking styles could be obtained in the model space [8, 9]. The present paper deals with the first step of transposing these methods to German expressive speech, namely, assessing the quality of HMM-based synthesis when applied to the German language. We are aware of two recent applications of HMM-based synthesis to German [10, 11]; however, these works have used limited training sets, and have therefore reached limited intelligibility. In contrast, we have used a speech database targeted at the development of unit-selection systems, which includes more than 3 hours of speech for each of four speakers, and enforces an optimal coverage of the German diphones. In addition, we present some preliminary results about the adaptation of the synthesis system to a very limited set of expressive football comments. This represents an exploratory step in anticipation of the recording of a larger and more diverse expressive speech database, that will better support expression-dependent model adaptation and the development of model-space methods for the explicit control of expressiveness in the synthesis of German.

While section 2 summarises the HMM-based synthesis technique, section 3 describes its application to the synthesis of German, with a comparison over distinct adaptation sets and various selections of context features. We comment the obtained results in section 4 and conclude about future work.

2. HMM-based speech synthesis

The publications about the key aspects of HMM-based speech synthesis are referenced on the homepage of the HTS open source software (`http://hts.ics.nitech.ac.jp/`). We only give a short summary in the following.

The HMMs underlying speech synthesis implement the same modelling logic as in speech recognition, namely, representing speech as a constrained sequence of random observations characterised by their second-order statistics. Significant differences from the HMMs used in speech recognition include:

• the explicit description of the pitch, by adding the log-scaled fundamental frequency (log-F0) and its 1^{st} & 2^{nd} order derivatives to the usual feature vector of Mel-Frequency Cepstrum Coefficients (MFCCs) which describes the spectral envelope;

• the use of Multi-Space Density functions, to accommodate a discrete voiced/unvoiced decision variable observed in conjunction with the continuous log-F0 values;

• the definition of so-called full-context models, which expand the n-phones with a richer set of context descriptors that go beyond the co-articulation effects, and which are related, e.g., to the lexical or syntactic levels of the training sentences. This entails a combinatorial increase of the number of context-dependent models, and a problem of sparsity of the training data available for each model. This problem is tackled by the application of standard tree-based state clustering techniques [12];

• a separate state duration model is trained for each context-dependent model on the basis of the state occupancies over the training set.

In speech recognition, the trained models are used as templates to be matched via the likelihoods of incoming observations. Conversely, for synthesis, a speech parameter generation algorithm is applied to emit some smooth sequences of synthetic MFCC and log-F0 features, in Maximum Likelihood accordance with a selected sequence of states.

3. Building HMM-based German voices

Practically speaking, our experiments rely on a modified version of the demonstration scripts delivered with the HTS 2.0 open-source synthesis software (`http://hts.ics.nitech.ac.jp/`). Our modifications cover the adaptation to the BITS and Bundesliga German databases, and the use of our own context features.

3.1. Training and adaptation data

An average German voice, denoted *world model* in relation to speaker recognition terminology, was trained by pooling the unit selection sets of the BITS German speech synthesis corpus [13]. This corpus contains 1683 sentences designed to have an optimal coverage of the German diphone space and spoken by each of 2 male and 2 female speakers. The recordings were made in a sound-proof, low echo room, at a 48KHz sampling rate/16bits resolution and using professional recording equipment. They were later down-sampled to 16KHz for our experiments. Of the original utterances, about 1500 sentences per speaker were compatible with our alignment and context feature extraction schemes, for a total of about 6000 sentences in the world model training set. As far as prosodic representation is concerned, the sentences are by a vast majority affirmations spoken in a matter-of-fact, read speaking style.

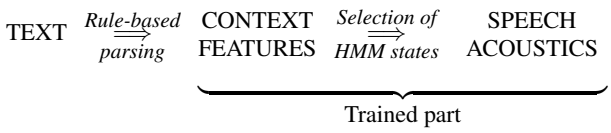
To produce four "standard" HMM-based German voices, the world model has been re-adapted to the individual BITS speakers, using their full set of 1500 sentences (about 3 hours of speech per speaker).

A locally recorded limited domain database, denoted "Bundesliga database", has also been used in a preliminary experiment of adapting the world model to a corpus presenting a limited phonetic coverage but a specific expressive speaking style. More specifically, the Bundesliga corpus contains speech from one male non-professional speaker uttering acted football announcements of two types: introductions, such as *"Und hier die Ergebnisse des [ersten—zweiten—etc.] Spieltags"* ("And here are the results of the [1st—2nd—etc.] round"), and results, such as *"[Club X] besiegt [Club Y] mit [Punktzahl]"* ("[Club X] beats [Club Y] with [score]"). 58 such sentences have been recorded in a neutral announcement style, and 52 in an excited announcement style encouraged by immersing the speaker in a stadium audio scene played through headphones. The excited style is characterised by a high vocal effort, high pitch level and range, steep and mostly falling intonation contours, and an increased speaking rate, whereas the neutral style is closer to the average characteristics of a neutral male voice. The energy profile of the excited voice has been equalised by the recording conditions, so it is not considered as a prosodic feature for the rendering of excitement. The recordings have been made in a sound treated room, using a microphone on a stand placed on the side of the mouth to avoid plops, and connected to a computer sound card recording data at a 16KHz sampling rate/16bits resolution. They have been manually labelled by a trained phonetician.

Two "football" voices, a neutral one and an excited one, have therefore been obtained by the independent adaptation of the world model to each of the small Bundesliga sets (about 5 minutes of speech per set).

3.2. Comparing various selections of context features

The full-context HMMs are trained to perform a mapping between a vector of context features and the corresponding speech acoustics:



The choice of the context features is crucial to the degree of perceived naturalness of the synthetic voice. For example, generating phrase breaks can only succeed when suitable descriptors are available, such as punctuation-related information. The context features can be of two types: (1) the features which are *implicitly* related to an acoustical realization, such as the punctuation or the word positions; (2) the features which are *explicitely* related to an acoustical realization, such as the phoneme sequence, the syllable stress, the phrase breaks or the ToBI tones, and which may be subjected to manual corrections in the labelling of the training data. Two training strategies are therefore conceivable: (a) training the HMMs from purely automatic text-to-feature labelling, which incurs the risk that if a significant number of pronunciations deviate from the predicted labels, then the HMMs will represent the context in a poor way; (b) training the HMMs over manually corrected labels, to enforce an accurate modelling of the context-dependent acoustics. If the manual corrections happen to be inconsistent with the automatic text-to-feature predictions, then some acoustical modelling capacity may be wasted in models which will never be requested to emit speech, or conversely the synthesis may systematically force the HMMs to generalise to unseen contexts. Therefore, in case (b), the modelling performances of the HMMs should be assessed independently of the text-to-feature accuracy, i.e., over an independent test set that would be calculated and corrected over human realizations. However, given the high dimension of the context feature vectors, setting a statistically significant test set apart from the training data would reduce the training set too much. The present work deals with case (b), but considers that it is sufficient to use sequences of context features issued from the training set to assess the acoustical modelling capabilities of the HMMs. Besides, the assessment of the overall TTS result assumes that the automatic text-to-features predictions correspond to plausible human realizations, and thus do not introduce artifacts in the assessment of the generalisation capabilities of the models. With these considerations in mind, our informal listening assessment uses sentences of three kinds: from sequences of context features seen in the

	Sys. 1	Sys. 2	Sys. 3
# context features	5	5 + 24	5 + 52
# fullcontext models	79 711	170 476	225 573
# questions	440	1 206	11 952
# tie states, MCep	3 132	3 254	3 304
# tie states, log F0	4 371	5 227	5 543
# tie pdfs, durations	712	799	801

Table 1: Dimensions of the three versions of the system.

training set, from text seen in the training set, and from unseen text (the classical German text "Die Buttergeschichte").

The test sentences have been compared over three voice modelling systems, corresponding to three different selections of context features. These features have been computed from the German text using the MARY speech synthesis platform (`http://mary.dfki.de/`), and they are listed in Table 2. **System 1** uses the description of the phonetic context only: the models are plain quintphones, coupled with an additional set of questions related to broad phonologic classes. **System 2** augments System 1 with a set of lexical and syntactic features which can be reliably predicted from the text. **System 3** augments System 2 with a set of features and questions related to the realization of Tones and Break Indices (ToBI). These features are expected to bring a better naturalness, but their prediction from text may be inaccurate, and may thus incur the above-mentioned generalisation problems. Table 1 summarises the dimensions of the three systems, in terms of the total number of context features, the resulting number of fullcontext models, the number of questions available for the building of the state clustering trees and the number of states after the state tying, which is performed independently for the Mel-Cepstrum and log-F0 acoustic subspaces. The phoneme duration is also modelled independently, by a set of tied 5-dimensional Gaussian pdfs (one dimension per state in the HMMs).

Phonetic features, Systems 1, 2 and 3
• **phoneme ID** (SAMPA set) plus some phonological features (vowel length/height/fronting/rounding, consonant type/place/voicing), propagated in a quintphone context (characteristics of the two preceding and two following phonemes)
Lexical and syntactic features, Systems 2 and 3
• **phoneme and syllable structure**: position-in-syl, position-type (single, final, initial, mid), syl-numsegs, segs-from-syl-{start,end}, word-numsyls, syls-from-word-{start,end} • **word related**: word-numsegs, segs-from-word-{start,end}, sentence-numwords, words-from-sentence-{start,end} • **punctuation related**: sentence-punc, {prev,next}-punctuation, words-{from-prev,to-next}-punctuation • **lexical stress**: syl-is-stressed, syls-{from-prev,to-next}-stressed • **part-of-speech tag** • **word unigram frequency class**, on a log scale from 1 to 10
ToBI and phrase related features, System 3 only
• **accents**: tobi-accent, {next,prev}-is-accented, {next,nextnext}-tobi-accent • **end tones**: tobi-endtone, phrase-endtone, prev-phrase-endtone, {next,nextnext}-tobi-endtone • **breaks**: syl-is-break, prev-syl-is-break • **syllabic locations across ToBI characteristics**: syl-is-accented, syls-{from-prev,to-next}-accent, accented-syls-from-phrase-{start,end} • **phrase structure**: sentence-numphrases, phrases-from-sentence-{start,end}, phrase-numsyls, syls-from-phrase-{start,end}, stressed-syls-from-phrase-{start,end}, phrase-numwords, words-from-phrase-{start,end}

Table 2: List of the context features delivered by the Mary system, with their affiliation to the compared systems.

4. Results

A trained phonetician (one of the authors) listened to the synthesised test sentences. Our informal observations are reproduced below, and can be compared to the audio files attached to the paper. They account for a preliminary exploration of the HMM synthesis capabilities in the context of German, in anticipation of a more precise focus on expressive speech modelling, which will involve better data and more formal listening assessments.

4.1. Global intelligibility and prediction of prosody

Despite the buzzy sound caused by the employed vocoding method, the generated speech is of a very stable quality and very high intelligibility. This is true for the *world model* voice and for all the adapted voices, with the exception of the excited voice (see below). In particular, German consonant clusters (e.g., the [nts-fr] of "ganz früh"–"very early"), which represent a specificity of the language, are rendered naturally (cf. `sys1-consclust.wav`).

The richness and appropriateness of the prosodic patterns depends, as expected, on the number and type of the considered context features. System 1, which uses only quintphones and derived phonological features, shows the flattest prosody (cf. `sys1-butter.wav`). A falling pattern can be observed on each phrase-final syllable; in addition, local pitch movements are present which apparently reflect micro-prosodic effects caused by the phonetic context. As expected from the limited context description, neither a global intonation contour nor a duration pattern reflecting the stress status of syllables are present. Nevertheless, the speech is clear and intelligible.

Systems 2 and 3 include a rich set of linguistic and prosodic context predictors, and thus show a considerably richer prosody. Word stress and phrase accent patterns are clearly perceptible and, in their vast majority, appropriate. The differences between the two systems are relatively small (cf. `sys2-butter.wav` and `sys3-butter.wav`). The prosody generated by system 2, which contains only "objective" context predictors such as punctuation, position in the sentence etc., does not sound "worse" than the prosody generated by system 3, which adds some explicit ToBI descriptions of the prosody.

However, system 3 generates a small number of errors that do not appear to the same extent in system 2, such as a consistently exaggerated high pitch and long duration on the function word "zu" in pre-final position (cf. `sys2-zu.wav` and `sys3-zu.wav`). It is conceivable that this is an example of overfitting: the number of predictor variables is so large that patterns in the training data can be modelled very closely, but they do not generalise to unseen data. More detailed investigations, notably of the decision trees predicting F0 and duration on the problematic instances of "zu", would be needed to clarify this point.

For the three systems, we compared the test sentences based on the manually corrected phoneme chain and (for system 3) ToBI prosody labels, with the sentences based on the fully automatic prediction of the features from text. Results show that the synthesis method is robust against this difference, in the sense that both types of features are rendered as speech of the same acoustic quality. Differences that can be perceived are on the level of phonetic identity, when the pronunciation observed in the BITS recordings differs from the pronunciation predicted by the TTS system. This is of particular interest with respect to post-lexical phonological processes such as Schwa elisions, which are frequent in German, even in careful speech such as the BITS recordings. Where a speaker did not realise a Schwa, this is reflected in the database labelling, and thus in the label-based synthesis result (cf. `sys2-lab.wav`); the TTS system, on the other hand, does not model such elision processes, and will thus always realise the Schwas (cf. `sys2-gen.wav`). This leads to a noticeable impression of hyper articulated speech, which could be attenuated by modelling some reduction rules.

4.2. Adaptation to specific voices

The specific voices created by means of adaptation capture the speaker specificities to a certain extent. The voices issued from the four BITS speakers show spectral characteristics that "resemble" the respective speakers, and the average pitch level is adapted to the speaker settings (compare `bits.wav` and `sys2-lab.wav`). However, the generated pitch range is considerably smaller than for the respective original speaker. This is a known problem related to the speech generation algorithm, which, in the version we had access to, does not enforce the global variance of the speech parameters over the adaptation set [14].

For the neutral Bundesliga voice, the global speaker characteristics are also captured reasonably well, including spectral properties, pitch level, and a slight Saarland dialectal colouring, and the voice is of stable quality (cf. `neutral.wav`). These performances are reached in spite of the very small and phonetically unbalanced amount of adaptation data. However, it sounds more "buzzy" than the BITS voices.

The excited Bundesliga voice, which is at the same time the smallest adaptation set and the most non-standard one, poses more problems. The global pitch level is adapted, but the spectral characteristics do not capture the full extent of the high vocal effort present in the original recordings (cf. `excited.wav`, `excited-foot.wav`). Instead, the voice sounds "squeaky", which may be related to the variance problem mentioned above. In addition, a wavering energy contour can be heard in some files, which was not the case in the *world model* voice or in any of the neutral adaptation voices. This suggests that adaptation across such wide expression-related acoustical differences requires a larger set of adaptation data.

5. Conclusions

The exposed results validate the use of HMMs as reliable models of German speech, in the sense that they are able to support the production of high-quality synthetic sentences, so far in the context of a neutral speaking style. It is verified that the choice of the context features that define the model set has an impact on the quality of the result. Besides, the voice adaptation paradigm that was developed for Japanese appears valid in the context of German. However, the adaptation to a small dataset of very expressive speech has reached a limited success: a larger database would be needed to investigate further the capacity of model adaptation to capture the expressivity in German. Recording such a database is part of the research plan promoted by the PAVOQUE project, which also includes the future development of more formal listening tests.

6. Acknowledgements

This work is supported by the DFG project PAVOQUE and by the EU project HUMAINE (IST-507422). We are grateful to the HTS and HTK development teams for making their software open source and publicly available.

7. References

[1] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Proc. Blizzard Challenge'06*, 2006.

[2] G. Baudoin and Y. Stylianou, "On the transformation of the speech spectrum for voice conversion," in *Proc. ICSLP'96*, Philadelphia, PA, USA, October 1996.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. Eurospeech'01*, 2001.

[5] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech'97*, 1997.

[6] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP'02*, Denver, Colorado, September 2002.

[7] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. Eurospeech'03*, Geneva, Switzerland, 2003, pp. 2461–2464.

[8] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *Proc. Speech Prosody 2004*, Nara, Japan, March 2004.

[9] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. ICSLP'04*, Jeju, Korea, 2004.

[10] C. Weiss, R. D. S. Maia, K. Tokuda, and W. Hess, "Low resource hmm-based speech synthesis applied to german," in *Proc. ESSP'05*, Prag, Czech Republic, 2005.

[11] C. Plahl, "Sprachsynthese mit hidden markov modellen," Master's thesis, Bielefeld University, 2005, (in German).

[12] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Workshop on Human Language Technology*, 1994.

[13] T. Ellbogen, A. Steffen, and F. Schiel, "The BITS speech synthesis corpus for German," in *Proc. LREC'04*, 2004.

[14] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. Interspeech'05*, 2005.