Investigating HMMs as a parametric model for expressive speech synthesis in German

Sacha Krstulović, Anna Hunecke and Marc Schröder

DFKI GmbH, Saarbrücken, Germany

firstname.lastname@dfki.de

ABSTRACT

The paper investigates the potential of HMM based synthesis to support the parameterisation of expressive speech in German. First, we review the assets of HMMs in the perspective of previous works in speech modelling and speech transformation. It is shown that HMMs define a flexible parametric model of the speech acoustics, which readily integrates several levels of speech modelling, such as distinct predictors for prosody and voice quality. HMM-based synthesis has also supported cross-speaker and cross-speaking style transformations with a good level of perceptual quality, albeit in other languages than German and over a limited range of styles. To try these considerations in our research framework, we have therefore performed a preliminary application of HMM technology to the synthesis of excited football announcements in German. It is shown that a highly intelligible voice can be obtained, but that the rendering of the prosodic and voice quality correlates of excitement could benefit from some improvement in well identified areas.

1. INTRODUCTION

The question of synthesising expressive speech may be interpreted as parameterising and manipulating the prosody and voice quality characteristics of speech in a synthesis system [13]. The goal of our work is to develop this consideration in the framework of the synthesis of German speech.

While formant-based systems offer some extended speech parameterisation capabilities, the perceptual quality of the produced speech is low. At the other end of the spectrum, the state-of-the-art unit-selection based synthesis offers good perceptual quality but is very rigidly linked to the contents of the underlying voice database. In the course of the past 10 years, Text-To-Speech (TTS) synthesis systems based on Hidden Markov Modelling (HMM) of the speech acoustics have emerged and have gradually reached a level of quality that makes them competitive with unit-selection systems [21]. While they have been successfully applied to speech transformations for the Japanese language, the present paper investigates their use as a parametric model of expressive speech in German.

Section 2 motivates the use of HMM-based systems for synthesising expressive speech. Section 3 describes a preliminary application of HMM technology to the synthesis of expressive football comments in German. The results are assessed in Section 4. Section 5 concludes on the validity of HMM as a parameterisation and synthesis tool for expressive speech in German.

2. HMMS FOR EXPRESSIVE SPEECH: MOTIVATIONS

The HMM-based synthesis framework integrates several methods inherited from various fields of speech modelling:

1. the modelling of the acoustics in the Cepstral domain, which has long been known to linearly decouple the glottal shaping effects from the vocal tract filtering effects [11], and which allows for a perceptually meaningful comparison of acoustical observations through some simple Euclidean or Mahalanobis distances [5];
2. Gaussian modelling, which has emerged over the years as the best supporting model for cross-speaker voice transformation [2, 14], and which allows for a mixture of supervised (e.g., phoneme-dependent) and unsupervised mapping in the same framework;
3. model adaptation, which has been widely deployed in speech recognition [17], and which extends the principle of cross-speaker or cross-channel Gaussian-based mapping to the HMM domain;
4. tree-based context clustering, used to implement the selection of speech units in unit-selection systems [3], as well as some context-dependent models of pitch and duration [12];
5. tree-based state tying [20], which aims at compensating for unseen speech units.

The above points can be assessed in the perspective of expression transformation: property 1 is important if one wishes to linearly de-couple the voice quality, related to varied glottal excitation modes, from the vocal tract filtering, more related to the phonetic contents; property 2 and 3 present Gaussian modelling and HMMs as a generic transformation tool for the speech acoustics, which could prove valid for cross-expression transformations; property 4 is directly related to the modelling of the prosodic characteristics of varied speaking styles; property 5 compensates for the curse of dimensionality which affects speech units when more expressive contexts are to be considered in a speech database.

Similarly to unit-selection, HMM-based synthesis relies on a data driven acoustic model, in the sense that the system learns a statistical model from an acoustic sample which covers the speech domain to be synthesised. This is in contrast to the methods for which a production model, based on expert knowledge, is available, such as formant-based or articulatory speech synthesis. Implementing some voice quality transformations in a speech production framework would imply, e.g., manipulating explicit measurements of glottal wave shapes, possibly based on a shape model or on a physical model; but one major challenge remains the automatic estimation of the production parameters from the speech data [1].

Another approach would be to try and implement an online conversion of the spectral envelopes of speech units in a unit-selection system. Indeed, a state-of-the-art unit conversion method would be supported by a Gaussian Mixture Model (GMM), an aspect which readily underlies the HMM paradigm. While unit-selection keeps track of a large set of speech units and makes only minimal parameterisation of these, the GMMs are able to reduce a whole set of speech units to a smaller set of model parameters such as variances, means and tree structures; hence, transformations operated via GMMs or HMMs entail a reduction of the number of parameters necessary to reach a good accuracy of the transformation.

From a perceptual point of view, an advantage of the HMM synthesisers over unit-selection is that their results are acoustically continuous by construction. Conversely, their main drawback is the vocoded quality of the result, related to artifical filter excitations by pulse trains or white noise; however, signs of improvement are visible in recent works, either from changing the excitation model [19] or from changing the acoustic feature space [9, 6].

Considering the above, we see HMM-based synthesis as a promising option for an application to the parameterisation of expressivity in speech synthesis. As a matter of fact, it has been recently applied to the modelling of speaking style variations in Japanese [18], and preliminary works towards a more explicit manipulation of style-related parameters are beginning to appear [10, 15]. However, these works have been applied in the framework of four particular speaking styles (reading, rough, joyful and sad), and for the Japanese language only. In our framework, a necessary preliminary step is therefore to assess the performance of HMM synthesis for the modelling of expressive speech in German.

3. EXPERIMENT: THE BUNDESLIGA GERMAN FOOTBALL VOICE

We carried out a first experiment to assess the properties of currently available HMM synthesis technology for the generation of expressive speech. The experiment implements the adaptation of a neutral German HMM voice to a limited set of highly expressive football (soccer) announcements from a single speaker. Practically speaking, it relies on a modified version of the demonstration scripts delivered as a complement to the HTS open-source synthesis software [7]. Our modifications cover the adaptation to the BITS and Bundesliga German databases, and the use of our own context features.

3.1. Training data

Training set for the average voice – The neutral German voice, denoted *world model* in relation to speaker recognition terminology, is an average voice model trained over the unit selection recordings of the BITS German speech synthesis corpus [4]. This corpus contains about 1500 sentences designed to have an optimal coverage of the German phonetic space, and spoken by each of 2 male and 2 female speakers for a total of about 6000 sentences. As far as prosodic representation is concerned, the sentences are by a vast majority affirmations spoken in a matter-of-fact, read speaking style.

Adaptation sets – In contrast, a locally recorded limited domain database, denoted "Bundesliga database", was used for the adaptation. This corpus presents a limited phonetic coverage but a specific expressive speaking style, corresponding to announcements in a football stadium. More specifically, it 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 speech rate, combined with a slight Saarland dialectal colouring and slightly slurred pronunciation.

The Bundesliga 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. Automatically assigned labels were manually adjusted by a trained phonetician.

For our synthesis experiment, we have performed the adaptation of the generic model over two sets, one with the 58 neutral sentences and one with the 52 excited sentences. In addition, the generic model has been re-adapted to the 1500 sentences of each of the BITS speakers to obtain four "standard" German HMM voices.

3.2. Context descriptors

The procedure to create HMM-based synthesis voices involves the definition of so-called full-context models, which characterise each phoneme in the database in terms of its phonetic and linguistic properties. The full-context models are clustered into acoustically similar models using decision trees, whose nodes test various properties of the context. At synthesis time, these decision trees are used to obtain the most appropriate HMM state sequence for any sequence of target context descriptors. The choice of context descriptors is crucial – for example, generating phrase breaks can only succeed when suitable descriptors are available, such as punctuation-related information.

For German, we computed a set of context descriptors issued on the one hand from the database labels and on the other hand from features automatically predicted from the text using our text-to-speech synthesis system MARY. The segmental phonetic context was defined using quintphones, together with the corresponding phonological features (vowel/consonant, consonant type/place, vowel height/place/rounding, etc.). Linguistic and prosody-related context features include part of speech, sentence punctuation, lexical stress, and word unigram frequency, as well as rule-based predictions of ToBI accents and phrases.

3.3. Synthesis of test sentences

After training the HMMs, a set of 80 test sentences was synthesised from a selection of prompts compiled from the training data and from unseen material (the classical German text "Die Buttergeschichte"). This selection ensures that for each voice, both "seen" and "unseen" contexts are synthesised, thus allowing to judge both the reproduction of speech from observed contexts and the capacity of the models to generalise to unseen context configurations. The "seen" test sentences were synthesised from the full context models based on the database labels; the "unseen" sentences were synthesised from text.

In addition, we have performed a low-level copy synthesis for the "seen" material, by feeding the Mel cepstrum and voicing+log-F0 parameters measured from the original recordings into the Mel Log-Spectral Approximation (MLSA) vocoder [8] that is normally used to generate the speech waveforms from the HMM-generated synthetic features. This is meant to assess the extent to which the vocoding method influences the realization of the speaking style in its own respect, independently of the HMM modelling performances.

4. RESULTS

The synthesised sentences were investigated by a trained phonetician (one of the authors). Observations are discussed in relation to audio examples which can be considered representative of the overall quality. These examples come as attachments to the paper, and are thus also subjected to the independent opinion of the reader.

4.1. General observations

Overall, it can be noted that HMM synthesis performs well for the German language. All voices are highly intelligible. The prosody predicted from symbolic context descriptors is mostly appropriate, with clearly perceptible phrase boundaries where indicated by symbolic context, level intonation at sentence-internal phrase boundaries, falling intonation at the end of a statement, and rising intonation at the end of a question. Lexical stress appears to be properly reflected in the duration patterns. Content words, for which symbolic ToBI accents are predicted, are usually more prominent than unaccented syllables, due to a combination of duration and pitch cues.

Suboptimal aspects can also be noted. Apart from the slightly "robotic" sound inherent to the artifical excitations currently employed in the MLSA vocoding technique, some limitations can be observed in the generated prosody (cf. audio file 1). Globally, the pitch range sounds compressed, giving a rather flat impression of intonation; occasionally, a function word is realised with too high a prominence, or a content word seems too short. Additional observations, explicited in the following section, suggest that this could possibly come from a wrong variance of the synthesized speech features. However, further research will be needed to determine the role that could likewise be played by the context features used as predictors, the quantity and nature of the training data, or the prediction method as such, in the observed limitations.

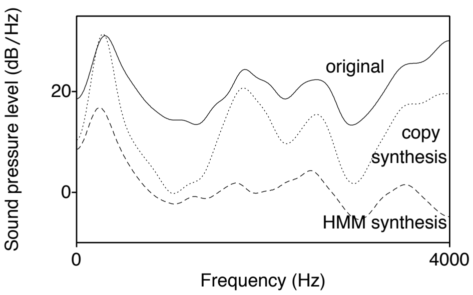

Figure 1: Cepstrally smoothed spectrum of [y:] from the team name "Duisburg", from audio files 6, 7 and 8.

4.2. Adaptation

Across the voices, some core speaker characteristics were realised. The spectral voice characteristics related to speaker identity appear to be preserved reasonably well, especially for the BITS voices (cf. audio files 2 and 3). Also, the pitch level of the training data is preserved.

Despite the small amount of training data available for the neutral and the excited Bundesliga adaptation sets, intelligibility is maintained at a high level in the resulting voices, and some voice characteristics are preserved: the average pitch level, traces of the dialectal colouring, and to a limited extent the voice identity are reproduced in the synthesised versions. This can be heard by comparing the original recordings (audio files 4 and 6) with the corresponding synthesised versions (audio files 5 and 7).

However, in the case of excited speech, the impression of excitement is not preserved after synthesis (compare audio file 7 with file 6). Even though the pitch level is the same, several other properties of the excited football announcements are not reproduced in the synthetic speech. Regarding prosody, the speaking rate and the pitch range are considerably smaller than in the original files; indeed, the average F0 standard deviation in the speech generated by the "excited" HMM voice is only 33% of the average F0 standard deviation in the original excited data.

On the level of voice quality, the impression of high vocal effort which can be perceived from the original recordings (audio file 6) is not preserved in the synthesised version. Instead, the voice rather sounds "squeaky" (audio file 7). Figure 1 shows that this perceptual difference is reflected in markedly different spectra: the clear peaks and valleys visible from the original are very much flattened in the spectrum generated from the HMMs. A plausible explanation would relate this flattening to the variances of the synthetic Mel cepstra, which are systematically and significantly smaller than the variances measured over the adaptation set (40% smaller standard deviations on average). This is a known problem of the speech parameter generation algorithm which underlies HMM synthesis, and a solution has recently been proposed to match the variances of the synthetic parameters with the global variance (GV) of the traning data [16].

Regarding the influence of the MLSA vocoding method on the perceived quality, it can be heard (audio file 8) that copy synthesis preserves the voice characteristics to a large extent. This is also reflected in Figure 1, which shows that the spectral envelope of copy-synthesised speech is very similar in shape to the original. This suggests that the change in voice quality owes more to the HMM predictions than to some limitations of the vocoding technique.

In a preliminary investigation of the relative contribution of prosody and of spectral properties to the perception of excitement, we have created a number of mixes from HMM-synthesised and copy-synthesised utterances: combining the Mel cepstra predicted by the HMMs with voicing and F0 information measured from the recordings, and vice versa; constraining the phone durations of HMM-generated speech to the original durations; expanding the F0 standard deviation of HMM-generated speech to the original F0 standard deviation; and expanding the standard deviation of each of the Mel frequency cepstral coefficients to the original standard deviation. The preliminary conclusion is that prosodic factors seem to play as important a role in conveying the excitement as the spectral characteristics, and that considerable improvement in expressivity could be obtained if a reliable way to correct the F0 and Mel-cepstrum variances was available.

5. CONCLUSIONS AND PERSPECTIVES

In this paper, we have pointed out a number of theoretical considerations indicating a high potential of the HMM-based approach for expressive speech synthesis. In a first practical experiment building on existing data and software resources, we have found that a neutral German voice modelled by HMMs can be adapted to a small set of excited football announcement recordings while still producing highly intelligible speech, but that the prosodic and voice quality features conveying the excitement are not yet captured adequately.

Some of the observed limitations appear more easy to overcome than others. Global prosodic settings such as the speaking rate and the pitch range can be enforced, either through post-processing of the generated parameters or through the use of normalised z-scores. Intonation contours and spectral properties specific to a given speaking style may be captured by voice adaptation if sufficient data is available, but more research is needed to determine an adequate dimensioning of the adaptation set. The most challenging aspect seems to be the preservation of the original voice quality. A number of approaches can be tried, including peak-enhancement [9], the GV method [16] or the replacement of the MLSA vocoding method with an inverse filtering method able to keep some affiliations with a codebook of natural excitations (e.g., Code-Excited Linear Prediction).

We are currently planning to record an expressive database with a larger amount of expression-specific adaptation data. Such a corpus will support the development of better expression-dependent models, which are a pre-requisite to the investigation of an explicit parameterisation of expressivity through model-space methods such as model interpolation or eigenvoices. The dimensions of the new corpus will also accommodate more formal listening assessments.

6. ACKNOWLEDGEMENTS

This work is supported by the DFG project PAVOQUE and the EU project HUMAINE (IST-507422). Thanks to the HTS and HTK teams for releasing their software!

7. REFERENCES

[1] Alku, P. 2003. Parameterisation methods of the glottal flow estimated by inverse filtering. *Proc. VOQUAL'03*, Geneva, 81–87.

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

[3] Black, A.W., P. Taylor, P. 1997. Automatically Clustering Similar Units for Unit Selection in Speech Synthesis. *Proc. Eurospeech '97*, Rhodes, Greece, 601–604.

[4] Ellbogen, T., Steffen, A., Schiel, F.. 2004. The BITS Speech Synthesis Corpus for German. *Proc. LREC 2004* Lisbon, Portugal, 2091–2094.

[5] Gray, A.H., Markel, J.D. 1976. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-24(5):380–391.

[6] Hemptinne, C. 2006. Integration of the harmonic plus noise model (HNM) into the hidden Markov model-based speech synthesis system (HTS). Master thesis, IDIAP Research Institute, June 2006.

[7] HMM-based Speech Synthesis System (HTS) http://hts.ics.nitech.ac.jp/ visited 5-Jan-07.

[8] Imai, S. 1983. Cepstral analysis synthesis on the mel frequency scale. *Proc. ICASSP'83*, Boston, 93–96.

[9] Ling, Z.H., Wu, Y.J., Wang, Y.P., Qin, L., Wang, R.W. 2006. USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method. *Proc. of the Blizzard Challenge 2006 workshop*, Pittsburgh, USA.

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

[11] Oppenheim, A. W., Schafer, R. W. 1968. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electracoustics*, AU-16(2):221–226.

[12] Pitrelli, J., Bakis, R., Eide, E., Fernandez, R., Hamza, W., Picheny, M. 2006. The IBM expressive text-to-speech synthesis system for American English. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1099–1108.

[13] Schröder, M. 2001. Emotional speech synthesis: A review. *Proc. Eurospeech'01*, Scandinavia.

[14] Stylianou, Y., Cappé, O., Moulines, E. 1998. Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech and Audio Processing*, 6(2):131–142.

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

[16] Toda, T., Tokuda, K. 2005. Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis. *Proc. Interspeech'05*, Lisboa, Portugal.

[17] Woodland, P. 2001. Speaker adaptation for continuous density HMMs: a review. *Proc. ITRW on Adaptation Methods for Speech Recognition*, Sophia Antipolis, 11–19.

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

[19] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. 2001. Mixed excitation for HMM-based speech synthesis. In *Proc. Eurospeech'01*, Scandinavia.

[20] Young, S., Odell, J., Woodland, P. 1994. Tree-based state tying for high accuracy acoustic modelling. *Proc. ARPA Workshop on Human Language Technology*, 307–312.

[21] Zen, H., Toda, T.,Tokuda, K. 2006. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. *Proc. of the Blizzard Challenge 2006 workshop*, Pittsburgh, USA.