IDIAP RESEARCH REPORT

PROGRESS REPORT OF A PROJECT IN VERY LOW BIT-RATE SPEECH CODING

Milos Cernak Philip N. Garner Petr Motlicek

Idiap-RR-08-2012

FEBRUARY 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

idiap RESEARCH INSTITUTE

Progress report of a project in very low bit-rate speech coding

Milos Cernak, Philip N. Garner, Petr Motlicek

February 29, 2012

Abstract

Background work in various levels of speech coding is reviewed, including unconstrained coding and recognition-synthesis approaches that assume the signal is speech. A pilot project in HMM-TTS based speech coding is then described, in which a comparison with harmonic plus noise modelling is also done. Results of the demonstration project including samples of speech under various transmission situations are presented in an accompanying web page. The report concludes by describing and enumerating the shortcomings of the demonstration system that define directions for future work.

This work is a deliverable for the armasuisse funded project “RECOD - Low bit-rate speech coding”

Contents

1 Introduction

Speech coding techniques can be split into three levels that vary in transmission rates from around 4 kbit/s to around 50 bit/s: MPEG-like coding (Section 1.1), and two recognition-synthesis based speech coding systems, HMM-TTS based coding (Section 1.2) and HNM coding (Section 1.3). Decreasing transmission rate towards Very Low Bit-Rate (VLBR) speech coding introduces language dependency and loses the personality of the speakers. The research challenge is to decrease the impact of VLBR system on these two factors. In the following text we summarise the aforementioned speech coding techniques.

1.1 MPEG-like coding

The state of the art in audio coding is well represented by the MPEG (moving picture experts group) standards. In MPEG-1 the ubiquitous MP3 was introduced that makes use of perceptual limitations of the human ear to encode arbitrary signals. In MPEG-2, AAC (advanced audio coding) replaced MP3, bringing in the likes of Huffman coding and the MDCT (modified discrete cosine transform).

The above codecs make no assumptions about the content of the signal. This changed in MPEG-4 (14496-3, 2009), which is more of a toolbox, where different codecs could be used for different purposes. In particular, if a signal is known a-priori to be speech, it is possible to use a speech codec. MPEG-4 includes two speech coders:

- Code-excited linear predictive (CELP) Schroeder and Atal (1985) and
- Harmonic Vector Excitation Coding (HVXC) Nishiguchi et al. (1997) and Nishiguchi (2006).

CELP operates at 4.0–16.0 kbit/s and HVXC constant bit-rate (CBR) on 2.0–4.0 kbit/s. Using a variable bit-rate (VBR) technique, HVXC can also operate at lower bit-rates, typically 1.2–1.7 kbit/s. In contrast to CBR, VBR technique vary the amount of output data per time segment. The bits available are used more flexibly to encode the speech with fewer bits in less demanding passages, and with more bits in difficult-to-encode passages. The average bit-rate in HVXC VBR is 1.5 kbit/s, and it has essentially the same quality as 2.0 kbit/s CBR. HVXC provides communications-quality to near-toll-quality speech in the 100–3800 Hz band at 8 kHz sampling rate.

The HVXC encoder has two types of excitation signals: one that changes rapidly, and one that changes slowly, in the time sequence of the power spectrum. For voiced segments, it uses the harmonic coding of the spectral magnitudes and it employs CELP for unvoiced segments. The HVXC encoder thus encodes vector quantised Line-Spectral-Pair (LSP) parameters, voiced/unvoiced decision, pitch, spectral envelope, and vector excitation coding shape and gain. The HVXC decoder uses encoded parameters for fast harmonic synthesis and a pitch/speed control algorithm. Besides very low bit-rate, time scale and speech modifications are key features of MPEG-4 HVXC speech coding.

Many of the above techniques are employed in a free speech codec known as speex[1].

1.2 HMM-TTS based coding

1.2.1 ASR and HMMs

Within the last two decades, automatic speech recognition (ASR) technology has almost completely converged around a single paradigm: the hidden Markov model (HMM) (Holmes and Holmes, 2001; Lee, 1989). It can be argued that the main reasons for this convergence are the ability of the HMM to learn from and respond to training material and the availability of efficient decoding algorithms. These in turn follow from a robust mathematical framework (dating back to at least Baum et al., 1970) that dictates a principled manner in which to collect training material and present it to the algorithm. The HMM framework is almost completely data-driven. That is, it responds automatically to data with little human interaction required. A minimal amount of phonetic and linguistic knowledge is required even for a new language.

[1] http://www.speex.org/

The mathematical framework has also spawned a portfolio of peripheral technologies such as adaptation (Lee and Gauvain, 1993; Leggetter and Woodland, 1994) that allow HMMs to be tuned to particular environments. Examples include adaptation to language, speaker, domain and background noise, as well as alternative signal processing techniques and training methods. In general the peripheral technologies advantageously share the HMMs' data driven capabilities. They allow, for example, tuning to a particular user after a few minutes. Practically, the adaptation technologies are very important as they can allow the performance of a general system to be pushed above a usability threshold.

1.2.2 TTS and HTS

Until recently, one could have said that the same time period had also seen a convergence in text-to-speech (TTS) technology (sometimes called speech synthesis). The technology had converged around a paradigm where HMMs were used only to annotate the training data; they were subsequently discarded. The actual synthesis was achieved by unit selection and overlap and add (OLA) techniques (Holmes and Holmes, 2001). Like HMM technology, unit selection responds automatically to training data. However, in contrast to HMM technology, unit selection lacks the portfolio of adaptation techniques that could allow improvement, tuning and personalisation.

In the last decade, however, a new TTS paradigm has emerged based on ASR technology. The technique, known as the HMM speech synthesis system (H-triple-S, or HTS), can be thought of as an inversion of an HMM that allows speech to be synthesised as well as recognised (Zen et al., 2007, 2009). This is a conceptually natural progression; an HMM is fundamentally a generative model. At the 2005 Blizzard challenge — an evaluation of TTS methods — to the surprise of the TTS community, an HTS system was rated by listeners more highly than all the unit selection systems (Zen et al., 2008). In subsequent years, the two paradigms have proved to be largely comparable, although this is in part because of the constraints of the Blizzard framework (Karaiskos et al., 2008). Given larger, well prepared databases, unit selection techniques still have the edge in terms of synthesis quality, and commercial entities do still favour them.

HTS has an at least one overwhelming advantage over unit selection: Because it is based on an ASR technique, it can benefit from the portfolio of peripheral technologies developed for HMMs in ASR. Further, because these technologies are primarily data driven, the benefits are produced in a completely automatic manner, without the need for skilled human intervention. One recent persuasive example of this is the voice adaptation of Yamagishi et al. (2009b); this method starts with HMMs trained on many speakers and uses HMM adaptation techniques drawn from speech recognition (Lee and Gauvain, 1993; Leggetter and Woodland, 1994), to adapt the models to a new speaker (of the same language and with the same accent). Such adaptation would be impossible with a unit selection based system.

1.2.3 The FP7 EMIME project and cross-lingual speaker adaptation

The FP7 EMIME project[2] (Wester et al., 2010; Kurimo et al., 2010) was a multi-lingual speech recognition and synthesis collaboration funded by the European union. The goal of EMIME was to produce a speech to speech translation framework characterised by the ability of the speech synthesiser to mimic the voice of the talker. For example, a speaker of Japanese would be able to hold a conversation with a speaker of English, where both speakers speak their native language. The Japanese speaker appears to his colleague to be speaking English, and vice-versa, even though both are actually hearing synthetic speech. Note that EMIME did not research translation technology. This was a deliberate omission, allowing the partners to concentrate on ASR and TTS. The EMIME final deliverable uses an off the shelf translator such as that provided by Google.

An important novelty of EMIME, compared to other S2ST research was that it brought the work of Yamagishi et al. (2009b, 2010), that applied the concepts of average voices and adaptation to TTS into S2ST research. The TTS adaptation allowed us to build a new voice using much smaller amounts of training data than previously required. A demonstration of the potential capability of this technology is displayed at the "Thousands of voices" web site[3] (hosted at Edinburgh university). The site demonstrates thousands of individual TTS voices created by the speaker adaptation; we can see how a central averaged

[2] http://www.emime.org

[3] http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map-new.html

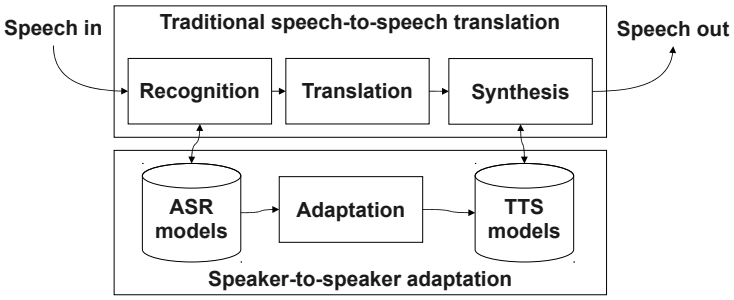

Figure 1: The EMIME concept

voice can be easily localised to any given region. The contribution of EMIME was to duplicate this functionality in a cross-lingual manner. That is, an average voice in, say, English, was adapted by a speaker speaking, say, Japanese. To this end, many cross-lingual speaker adaptation (CLSA) techniques have been developed (Wester et al., 2010). This is underpinned by the fact that both ASR and TTS are speaker adaptive thanks to HMM speaker adaptation technology.

CLSA was first considered by Wu et al. (2009). Within EMIME, Liang and Dines (2010) found that when the model order of the mapping is increased, the performance of intra-lingual speaker adaptation increased, but that of cross-lingual adaptation decreased because of the mixing of speaker and language transformations. Further work at Idiap has shown that use of bilingual data (where one speaker utters data in two languages) can alleviate this problem. Another route to cross-lingual adaptation is via a fundamentally language independent technique such as the VTLN of Saheer et al. (2010a).

The EMIME project brought together partners from universities at Edinburgh, Cambridge, Helsinki and Nagoya, together with Idiap and Nokia. The languages considered within EMIME largely reflected the local languages of the partner institutions: English, Finnish, Mandarin Chinese and Japanese. These also represented very different languages, able to stretch the capabilities of the system to its limits. Notably, the EMIME languages did not include any of the native Swiss languages.

1.3 HNM based coding

1.3.1 HNM analysis/synthesis

It is desired to find such a synthesis algorithm that would ensure a smoothed transitions from one acoustic (phonetic) unit to the following one, while preserving reasonably high subjective qualities of the resulting speech. Another aspect in choosing right synthesis technique is the possibility to divide split the encoding parameters into two parts: (1) parameters related to the prosody (i.e., the excitation signal), and (2) parameters related to speech production (i.e., cepstral or LPC coefficients, etc.).

There are various methods of representation and concatenation of acoustic units. Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) performs a pitch synchronous analysis and synthesis of speech (Dutoit, 1997). The concatenation in TD-PSOLA is not a trivial task due to its non-parametric structure even though the modification of prosody is easy. Multi-Band Re-synthesis Overlap-Add (MBROLA) attempts to overcome concatenation problems in the time-domain by re-synthesising voiced parts of the speech database with constant phase and constant pitch (Charpentier and Stella, 1986). During synthesis, speech frames are linearly smoothed between pitch periods at unit boundaries. Other possible models are based on sinusoidal approximation, Linear Prediction (LP), etc.

In this work, we have decided to apply Harmonic-Noise-Model (HNM) synthesis, which largely outperforms LP based synthesis techniques, and by comparison to other previously mentioned techniques, the prosody modification is quite simple due to the straightforward separation of the prosodic and speech-production parameters. HNM synthesis for very low bit-rate speech coding was investigated by Motlicek (2003). It is based on HNM modelling originally proposed by Stylianou (2006). The HNM is built on

Description	Number (*of floats*)
Gain in Harmonic part	1
Gain in Noise part	1
LSFs of Harmonic part	16
LSFs of Noise part	12
Pitch	1

Table 1: Structure of frames parametrised by HNM.

following representation of an input signal $x(n)$:

$$x(n) = \underbrace{\sum_{k=1}^{P} \alpha_k \cos(2\pi f_k n - \phi_k)}_{\text{Harmonics}} + \underbrace{b(n)}_{\text{Noise}}, \tag{1}$$

where P is the number of harmonics, α_k are the amplitudes, f_k the multiples of pitch and ϕ_k phases of harmonic part. $b(n)$ expresses components of noise.

HNM is composed of several blocks:

- Voiced/Unvoiced segment detection: In this block, the input frames are split into two categories, according to the voicing information they are carrying.
- Estimation of the fundamental frequency: In case of voiced segments, the pitch information needs to be appropriately estimated.
- Modelling of the segments: The segments detected as voiced are modelled by both Harmonic and Noise model, whereas unvoiced segments are modelled only by noise model.

In case of applying the Harmonic model, the estimation of amplitude and phase frequency coefficients is performed. The exact formula applied (slight modification of Eq. 1) is given by the following equation:

$$x(n) = \sum_{k=1}^{L} a_k \cos(2\pi f_k n) + b_k \sin(2\pi f_k n) + b(n), \tag{2}$$

where x represents N samples (one frame) of analysing speech signal.

In the case of the noise model, the model is applied on the signal obtained by subtracting the signal which is generated from the previously estimated harmonics and the original (input) signal. Its spectrum is modelled by LP (all-pole) filter of 12th order.

1.3.2 HNM coding

The way to exploit HNM synthesis for speech coding is quite straightforward. First, the input speech is classified into voiced/unvoiced segments. In the case of voiced speech, the parameters of the harmonics model are extracted together with the fundamental frequency. Further, LP parameters describing the noise model are extracted too. In the case of unvoiced segments, only the noise model parameters (LP coefficients) need to be extracted. In fact, parameters of the harmonic model are transformed into LP coefficients (i.e., an all-pole filter as is case in noise modelling).

In the case of speech coding, all the hitherto mentioned parameters need to transmitted from the encoder to the decoder. Such a system needs to deal with quantising the transmitting coefficients (as well as channel coding problems). For this reason, Line-Spectrum-Frequency (LSF) representation is chosen. Table 1 shows a detailed description of the transmitted parameters in HNM coding.

Note that these parameters are not transmitted directly in HNM speech coding. A detailed description of the coding system that can achieve around 200 bits/s transmission rate is introduced in Sec. 2.2.

2 Pilot project

This section describes in detail the experiments performed using recognition-synthesis based speech coding. More specifically, two different types of approaches were chosen. In the first one, Harmonic/Noise Model (HNM) analysis/synthesis was exploited and incorporated into a data-driven framework which can eventually perform speech coding on very low bit-rates. In the second approach, speech coding with very low bit-rates is achieved by integration of ASR and TTS, where a sequence of linguistic symbols (e.g., phonemes) or artificial units (e.g., coded speech chunks), is transmitted instead of compressed audio signal.

In this pilot project we investigated the both systems (described in Sections 2.1 and 2.2) that are partially complementary around two factors: multi-linguality and keeping the personality of the speakers. An audio and video demonstration of the pilot project, forming a key part of this report, is available at URL:

http://www.idiap.ch/project/armasuisse/recod/index.html

2.1 HMM-TTS based coding

The HMM-TTS based speech coding uses HMM speech synthesis (cf. section 1.2.2). The HMM approach has all the advantages of statistical rigour and peripheral technologies as for ASR. For instance, distinct voices have been produced from small amounts of adaptation data. Although the HMM and HTS paradigms unify the general theory of ASR and TTS, and there is still a significant practical gap between the two approaches, they can be integrated into an elegant solution of very low bit-rate speech coding. Fig. 2 shows the basic architecture that shares also EMIME concept.

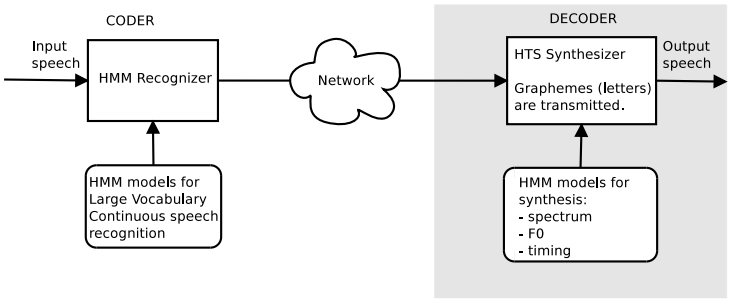

Figure 2: The HMM coder based on EMIME concept

Speaker individuality is by default completely discharged, but model adaptation techniques allow it to be kept partially or almost perfectly. Voice adaptation in HTS starts with HMMs trained on many speakers (HTS average training described by Yamagishi and Kobayashi (2007)) and uses HMM adaptation techniques drawn from speech recognition, to adapt the models to a new speaker (of the same language and with the same accent). Two adaptation were investigated:

1. The Vocal Tract Length Normalisation (VTLN) adaptation of Saheer et al. (2010a). A single VTLN transform for a speaker is calculated, and applied in a simple way, transforming the generated feature files. The target α_{target} is calculated (see Eq. 3) using the second term in the transformation matrix, α_{vtln}, and the synthesized spectrum is converted to new warping.

$$\alpha_{\text{target}} = (\alpha_{\text{original}} - \alpha_{\text{vtln}})/(1 - \alpha_{\text{original}} \times \alpha_{\text{vtln}}) \tag{3}$$

   Here, only α_{vtln} is transmitted (a single floating point number per sentence or several sentences), and it increases transmission rate negligibly. However, the adaptation performance is rather weak. Results are observable only when voices of different genders are adapted.

2. Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR) adaptation (Yamagishi et al. (2009a)) performs much better, but estimated bit-rates are much higher. We calculated CSMAPLR from the last 25 utterances spoken by the speaker; examples are on the audio demonstration page.

While speaker individuality can be adapted, cross-lingual speaker adaptation belongs rather to ongoing research topics, as introduced in Sec. 1.2.3.

A video demonstration of the proposed speech coding approach is available on the project page. The video demonstrates the baseline system that operates on the lowest limits around 50 bps (see Sec. 2.3 for more details). Only the recognised phonemes are transmitted through the communication channel - a TCP/IP connection. No adaptation is used, and so the voice of the synthesized speech is the same for all speakers. The baseline is thus a speaker dependent system. The upper terminal runs the HTS server (decoder) and the ASR client (encoder) runs in the lower terminal. The ASR module — Juicer described by Moore et al. (2006) — listens to the input microphone and at the end of each utterance, the sequence of phonemes (letters) are transmitted to HTS. The HTS module converts the transmitted sequence of letters back to speech.

2.2 HNM based coding

2.2.1 Detailed description

The current version of HNM based VLBR speech coding system employs several pre-processing and post-processing blocks:

- Analysis/Synthesis - The input speech segmented into "short-term" frames are passed to the block of HNM analysis. As a result, parameters reflecting the input speech frame are extracted (see Section 1.3.1). On the decoder side, a block of synthesis, generating the speech from extracted HNM parameters, is employed. HNM works in a similar way as traditional LPC, i.e., LPC parameters describing vocal tract/speech production system represented by an all-pole system, which is excited by an excitation signal on the decoder side. In case of HNM, more proper modelling of "voiced" speech frames is performed.
- Extraction of speech segments - Instead of using the speech frames (i.e., approximately 30 ms long speech segments extracted at the encoder side every 10 ms (as it is a case in traditional LPC analysis/synthesis), the VLBR speech encored/decoder uses much longer speech segments (rather called speech units). The speech units are automatically generated and are of arbitrary lengths. In order to generate such the speech units, block of "Temporal Decomposition (TD)" is used, described by Bimbot et al. (1988).
- Quantisation of speech units - In order to unify the speech units extracted in the previous block of TD, and to be able to represent these segments on a symbolic level (by an ID), the speech units are clustered using Vector Quantisation (VQ). The VQ training performs traditional K-means algorithm, based on Linde-Buzo-Gray (LBG) technique, where a codebook from previous iteration is always split into two sub-codebooks.
- Hidden Markov Models (HMMs) - HMMs are widely used in acoustic modelling for automatic speech recognition to classify the input speech into "phonetic" classes. In this work, HMMs are used in a similar manner. First, HMMs are used in the training process on top of VQ for iterative refinement of initially derived speech units. Then, HMMs are used in the task of encoding. In other words, trained HMMs are employed to classify the input speech into set of N classes, where N is equal to the number of clusters generated by VQ.
- Representatives - In order to be able to decode the speech initially classified by HMM system, the decoder side needs to share a dictionary of the speech unit representatives with the encoder side. These representatives (i.e., LPC or HNM parameters describing speech units (clustered into N clusters)) are generated by random selection (but satisfying initially selected criteria such as minimum length of the representative, ...) of speech units.
- Process of encoding/Decoding - During encoding, the input speech is first processed using TD+VQ or directly classified by well-trained HMM classifier. The output given in a form of a sequence of IDs (where each ID is representing one HMM class or VQ cluster) is transmitted from the encoder to the decoder. On the decoder side, the sequence of IDs is replaced by a sequence of representatives selected from the dictionary. Such a new sequence (in our case of HNM parameters) is synthesized (transformed into the resulting speech signal) using block of HNM synthesis.

- Additional prosodic parameters - Previously described steps are able to generate a synthesized speech where the parameters related to th production system (i.e., vocal tract) are transmitted on very low bit-rates. However, the information about the prosody (pitch, energy, temporal duration) is not transmitted and therefore not re-produced on the decoder side. The prosody information can be:
  - randomly generated (the final bit-rate will not be increased, however the quality of the resulting speech will be poor).
  - fully transmitted from the encoder to the decoder (very good quality of the speech, but high increase of the bit-rate).
  - partially transmitted (only the most important parameters are transmitted from encoder to decoder such as information about voicing, fundamental frequency and length of the synthesized speech units).

2.2.2 Overall description

The HNM based speech coding uses HNM synthesis that performs a speaker-dependent unit selection on the encoder side and an overlap-and-add (OLA) post-processing on the decoder side. Whilst unit selection also responds well to training data, as in case of ASR, adaptation to new speakers (voices) is more difficult.

HNM speech coding, graphically illustrated in Fig. 3, is based on Harmonic plus Noise Modelling of speech chunks, called also units (or segments). Each unit can be represented by several particular examples (developed during the training). In the audio demonstration page, the models of $N = 64$ units were trained from an English broadcast speech database. On the encoder side, the input speech is transcribed using an ASR module into a sequence of units (together with an additional information about the selection of the best example and the timing information). The decoder takes the unit sequence (while sharing the database of unit examples with encoder), and concatenates selected examples using OLA technique to produce the output speech.

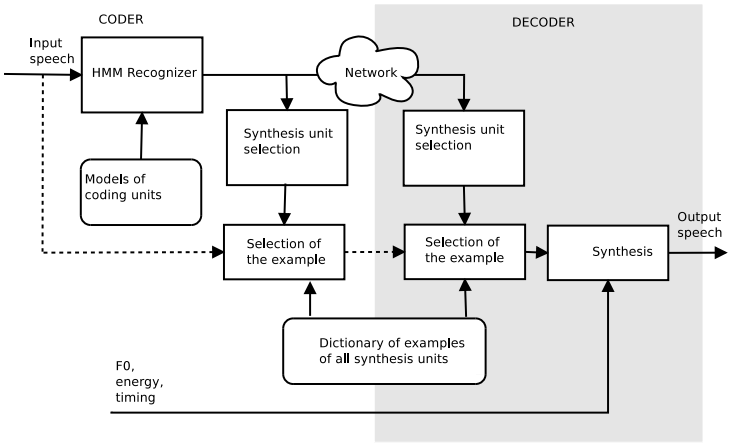

Figure 3: The HNM coder

Prosody information (i.e, timing information, speaker individuality and the like) is preserved on the encoder side. The full system is trained only on an English database. Due to that, for other languages such as French and German, the quality of synthesized speech pronounced in these languages is significantly worse. Nevertheless, this could be optimised as unit training and selection is based on a completely data-driven approach (and thus language independent).

Overall, the HNM coding system offers a language independent and speaker dependent approach. The quality can be significantly improved by accommodating current state-of-the-art speech synthesis techniques, as those mentioned in 2.1.

2.3 Transmission rates

The HMM-TTS based speech coding achieves a theoretical low limit, as only phonemes/graphemes are transmitted, 8-9 phonemes per second coded with 6 bits = ~50 bps. Transmitting just VTLN alpha factors keep low bit-rate, however the adaptation takes into account only the principal information (i.e., about gender). Nevertheless, to acquire language and speaker independence in the coding systems, additional information needs to be transmitted, increasing the bit-rate to around 200-300 bps.

The HNM based speech coding transmits roughly the same information, plus additional time-alignment information and prosody contour. The estimated transmission rate is around 200 bps.

3 Conclusions and recommendations

We have shown that low bit-rate speech coding is possible using technology that is currently available. A demonstrator has been constructed that joins together the component parts, allowing comparison of two distinct but closely related coding schemes.

Although the system works, the purpose of such a pilot project is to define future work. It is therefore appropriate to draw attention to the limitations of the current system, which serve to define the recommendations for future work:

1. The recognition-synthesis system is monolingual. To remedy this for a Swiss environment, we recommend that multilinguality be investigated for both the recognition and synthesis. Research would aim to find, for instance, whether it is better to merge several languages into one system or to incorporate language identification indicating one of several monolingual systems.
2. One approach to multilinguality could be to build an HMM-TTS system that is entirely phonetic. Such a system would suffer in terms of naturalness, but would gain in terms of vocabulary and language independence.
3. Speaker and cross-lingual adaptation in the recognition-synthesis system produces huge linear transformation matrices that are not possible to transmit in VLBR speech coding scheme. The linear transformation must be somehow approximated or alternative transforms must be used, such as VTLN that we tried.
4. The HNM approach is multi-lingual, however our demonstration showed that if the system was trained using monolingual speech database (English in our case), the quality of other languages is rather low (at least for Valaisan German and French). Multilinguality has to be investigated again here as well.
5. The ASR is currently driven by a VAD. For smooth communication, the ASR should be able to run in a true continuous mode.
6. HTS synthesis is a computationally expensive process and introduces unacceptable delay into real-time communication. Streaming HTS (Astrinaki et al., 2011) could be investigated for that.
7. The demonstration page also presents re-synthesis examples, where the quality of vocoders can be observed. It seems that the state-of-the-art speech synthesis technique HTS using the STRAIGHT vocoder of Kawahara (2006) outperforms the HNM vocoder. We propose further investigating vocoder selection, and also exchanging the vocoders between the two systems; for example using a STRAIGHT-like vocoder where HNM encoding/decoding was used earlier and vice-versa.
8. Both lower transmission rate speech coders in the pilot project operate with a set of up to 64 units (to use only 6 bits for encoding of unit labels). Units could be phonemes, as in the recognition-synthesis system, or automatically selected and quantised speech chunks as in the HNM system. It is important to optimise the unit selection system considering perhaps other state-of-the art classifiers instead of HMM. Increasing the number of units to 128 may be necessary for a multilingual system.
9. The HNM system does not transmit prosody. An investigation as introduced in conclusion of Section 2.2.1 must be done to remedy this.

4 Acknowledgements

All the work in this report was funded by armasuisse
http://www.ar.admin.ch/internet/armasuisse/en/home.html

References

ISO/IEC 14496-3. *Information technology Coding of audio-visual objects, Part 3: Audio*. ISO/IEC, Geneva, Switzerland, 4th edition 2009. URL http://webstore.iec.ch/preview/info_isoiec14496-3%7Bed4.0%7Den.pdf.

M. Astrinaki, O. Babacab, N. D'Alessandro, and T. Dutoit. sHTS: A Streaming Architecture for Statistical Parametric Speech Synthesis. In *First International Workshop on Performative Speech and Singing Synthesis*, 15 March 2011.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

F. Bimbot, G. Chollet, P. Deleglise, and C. Montacie. Temporal decomposition and acoustic-phonetic decod-ing of speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 445–448, 1988.

F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveform concatenation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1986.

T. Dutoit. An introduction to text-to-speech synthesis. In *Kluwer, Dordrecht, The Netherlands*, 1997.

John Holmes and Wendy J. Holmes. *Speech Synthesis and Recognition*. Taylor & Francis, 29 West 35th Street, New York, NY 10001, 2nd edition, 2001.

Vasilis Karaiskos, Simon King, Robert A. J. Clark, and Catherine Mayo. The Blizzard Challenge 2008. In *Proceedings of the Blizzard Challenge 2008*, Brisbane, Australia, September 2008.

Hideki Kawahara. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353, 2006. doi: 10.1250/ast.27.349. URL http://dx.doi.org/10.1250/ast.27.349.

Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichiro Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, and Junichi Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48–53, Uppsala, Sweden, July 2010.

Chin-Hui Lee and Jean-Luc Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, Minnesota, USA, April 1993.

Kai-Fu Lee. *Automatic Speech Recognition. The Development of the SPHINX System*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers, 1989.

C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England, June 1994.

Hui Liang and John Dines. An analysis of language mismatch in hmm state mapping-based cross-lingual speaker adaptation. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.

Darren Moore, John Dines, Mathew Magimai.-Doss, Jithendra Vepa, Octavian Cheng, and Thomas Hain. Juicer: A weighted finite-state transducer speech decoder. In *3rd Joint Workshop on Multimodal Interaction and Related Machine LEarning Algorithms MLMI'06*, 0 2006. IDIAP-RR 06-21.

Petr Motlicek. *Modeling of Spectra and Temporal Trajectories in Speech Processing*. PhD thesis, 2003. URL http://www.fit.vutbr.cz/research/view_pub.php?id=7281.

L. Nishiguchi, K. Iijima, and J. Matsumoto. Harmonic vector excitation coding of speech at 2.0 kbps. In *Speech Coding For Telecommunications Proceeding, 1997, 1997 IEEE Workshop on*, pages 39–40. IEEE, September 1997. ISBN 0-7803-4073-6. doi: 10.1109/SCFT.1997.623885. URL http://dx.doi.org/10.1109/SCFT.1997.623885.

Masayuki Nishiguchi. Harmonic vector excitation coding of speech. *Acoustical Science and Technology*, 27 (6):375–383, 2006. doi: 10.1250/ast.27.375. URL http://dx.doi.org/10.1250/ast.27.375.

Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang. Implementation of VTLN for statistical speech synthesis. In *Proceedings of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, September 2010a.

Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang. Implementation of vtln for statistical speech synthesis. In *Proceedings of ISCA Speech Synthesis Workshop*, number Idiap-RR-32-2010, 9 2010b.

M. Schroeder and B. Atal. Code-excited linear prediction(CELP): High-quality speech at very low bit rates. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, volume 10, pages 937–940. IEEE, April 1985. doi: 10.1109/ICASSP.1985.1168147. URL http://dx.doi.org/10.1109/ICASSP.1985.1168147.

I. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole National Supérieure des Télécommunications, 2006.

Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, Lakshmi Saheer, Simon King, Keiichiro Oura, Philip N. Garner, William Byrne, Yong Guan, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, and Junichi Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, September 2010.

Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 528–531, Brighton, U.K., September 2009.

J. Yamagishi and T. Kobayashi. Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training. *IEICE Trans. Information and Systems*, E90-D(2):533–543, February 2007.

J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm . *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):66–83, January 2009a. ISSN 1558-7916. doi: 10.1109/TASL.2008.2006647. URL http://dx.doi.org/10.1109/TASL.2008.2006647.

Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6), August 2009b.

Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Yong Guan, Rile Hu, Keiichiro Oura, Yi-Jian Wu, Keiichi Tokuda, Reima Karhila, and Mikko Kurimo. Thousands of voices for HMM-based speech synthesis–analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):984–1004, July 2010.

Heiga Zen, Keiichi Tokuda, and Tadashi Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, 21(1):153–173, January 2007.

Heiga Zen, Tomoki Toda, and Keiichi Tokuda. The Nitech-NAIST HMM based speech synthesis system for the blizzard challenge 2006. *IEICE Transactions on Information and Systems*, E91-D(6):1764–1773, June 2008.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1154, November 2009.