Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Audio Re-Synthesis based on Waveform Lookup Tables

Sebastian Heise[1], Michael Hlatky[1], and Jörn Loviscach[2]

[1] Accessive Tools GmbH, 28195 Bremen, Germany
{heise, hlatky}@accessive-tools.com

[2] Fachhochschule Bielefeld (University of Applied Sciences), 33602 Bielefeld, Germany
joern.loviscach@fh-bielefeld.de

ABSTRACT

Transmitting speech signals at optimum quality over a weak narrowband network requires audio codecs that must not only be robust to packet loss and operate at low latency, but also offer a very low bit rate and maintain the original sound of the coded signal. Advanced speech codecs for real-time communication based on code-excited linear prediction provide bandwidths as low as 2 kbit/s. We propose a new coding approach that promises even lower bitrates through a synthesis approach not based on the source-filter model, but merely on a lookup table of audio waveform snippets and their corresponding Mel-Frequency Cepstral Coefficients (MFCC). The encoder performs a nearest-neighbor search for the MFCC features of each incoming audio frame against the lookup table. This process is heavily sped up by building a multi-dimensional search tree of the MFCC-features. In a speech coding application, for each audio frame, only the index of the nearest neighbor in the lookup table would need to be transmitted. The encoder synthesizes the audio signal from the audio waveform snippets corresponding to the transmitted indices.

1. INTRODUCTION

Until today, the available bandwidth on IP networks has been constantly growing. This fueled the development of high-quality real-time voice communication applications such as Skype, iChat or Google Talk. The audio quality provided by such programs easily surpasses standard telephones, as long as sufficient network bandwidth is available. On weak, narrowband networks however, packet loss, jitter, the limited bandwidth and high transmission latency substantially degrade the perceived audio quality as well as the intelligibility of the transmitted speech signal. To improve on this, audio codecs have been optimized to work with low bandwidths, to be tolerant even to substantial packet loss and to introduce only a very small additional delay through their encoding and decoding processes. However, there seems to be a minimum bandwidth barrier for today's

speech codecs with recognizable, natural voice characteristics—around 2 kbit/s—which has not yet been significantly under-run. In this paper, we introduce a new approach for an audio codec that seems promising to break the 2 kbit/s barrier while maintaining the natural sound characteristics of the original voice.

Our approach works similar to many feature-based music similarity search engines: The incoming audio stream is analyzed on a basis of windowed frames with an MFCC-based feature extraction. The extracted feature vector of each frame is compared to a collection of feature vectors from a pre-generated lookup table holding the sinusoidal components of selected frames of audio data through a nearest neighbor search. Only the index of the nearest neighbor is transmitted. The decoder reads the list of indices, and re-synthesizes the coded audio using the corresponding waveforms.

2. RELATED WORK

2.1. Speech Coding

Different methods of coding speech have been introduced over the years, which can be characterized in three classes: Waveform coders, source coders and hybrid coders.

Waveform coders, who work directly in time domain, try to reproduce the digital representation of the analog speech signal as identical as possible to the original signal. The G.711 [1] standard, a waveform coder introduced in 1972, which is used in digital telephone networks, describes two waveform companding methods, the μ-law algorithm used in North America and Japan, and the A-law algorithm used in Europe and for international connections. The G.711 codec operates at a bandwidth of 64 kbit/s.

Military applications demanded lower bandwidths to allow heavy protection and encryption. The naturalness of the transmitted speech could be neglected here, as only a high intelligibility at low bit rates was necessary. Therefore source coders that can reproduce intelligible speech signals with substantially lower bit rates were employed. Source coders for speech signals reproduce a digital signal by using a model of how the source was generated (in the case of speech an excitation signal produced by the lungs and the glottis, which is filtered by the vocal tract) and extract by the signal being coded the parameters of the model, which are transmitted to the receiver. Such codecs can operate at bit rates as low as 100 bit/s [2]. Speech source coders, however, usually fail to produce naturally sounding speech output. Yet they may enable limited speaker recognisability by methods such as codebook adaption or applying adaptive Hidden Markov Models (HMM) [3].

Hybrid speech coders, as for instance analysis-by-synthesis coders use the same linear prediction model of the vocal filter tract as found in many speech source coders. However, instead of using only simple state-based models to find the necessary input to the filter, they chose the excitation signal by attempting to match the original and the reconstructed speech signal. Hybrid speech coders usually can produce naturally sounding output while maintaining reasonably low bandwidths.

With the move of digital telephony to cellular networks, codecs that cater for lower bandwidths for the transmitted signals, however, still produce natural sounding speech, were required. Hybrid speech coders as for instance the Regular-Pulse Excited Linear Predictive Codec (LPC-RPE) [4] standardized in 1991 and employed in the GSM radio network for instance operates at a bandwidth of 13 kbit/s.

Voice over IP (VoIP) scenarios, which may require even lower bandwidths than transmission over the GSM network, often use a hybrid codec of the Code-Excited Linear Prediction (CELP) [5] family. Their underlying principles were introduced in 1985. CELP-based coders, as for instance Speex [6] provide intelligible, naturally sounding speech output on bitrates as low as 2 kbit/s [6].

2.2. Automated Speech Recognition

Alongside the improvement of the digital transmission of speech, methods for automated recognition of spoken words were developed. Already in 1952, Davies et al. proposed a system for the automated recognition of spoken digits [7]. Modern speech recognition systems usually employ features such as Mel-Frequency Cepstral Coefficients (MFCC) derived from the audio signal alongside with statistical acoustic and language modeling employing for instance HMMs. MFCCs as acoustic features for speech recognition were first introduced in 1976 by Mermelstein [9].

As the quality degradation introduced by linear predictive speech coding sufficiently constrains speech recognition accuracy, Lee at al. [10] introduced a MFCC-

based CELP speech coder that even provides improved speech quality.

2.3. Audio Similarity Analysis

In the field of content-based retrieval of music and audio, MFCCs increasingly find applications in genre classification or as similarity measure [11]. In 1997, Foote proposed a system that uses a tree-based nearest neighbor search on MFCCs for content-based audio file retrieval [12].

2.4. Speech Synthesis

Beyond model-based audio codecs, naturally sounding speech synthesis is also applied in Text-To-Speech (TTS) systems. Already in 1953, Fant introduced the formant-based synthesis of vowels [13]. His findings led to the development of the source-filter model of human speech production, which is employed in many of today's speech codecs. The source-filter model is based on the assumption that speech is produced by the lungs generating pressure that is released through the glottis and then filtered by the vocal tract. The glottal excitation waveform is often modeled as a periodic pulse train for voiced and white noise for unvoiced sounds. The vocal tract is usually modeled as an all-pole filter which is capable of modeling the four formant resonant peaks produced by the human voice tract.

Another possibility to represent the glottal excitation waveform is as a sum of sinusoidal functions. This model was proposed by McAulay and Quatieri [14] in 1986.

Recently, attempts have been made to synthesize speech directly from ASR features such as MFCCs [15]. These methods lead to an intelligible estimate of the speech signal. However, the speech quality is poor when the reconstruction is solely based on the MFCC data, as the descriptor seems to not carry sufficient information about the pitch spectrum.

3. IMPLEMENTATION

The approach for a speech codec we propose in this paper is based on two assumptions: first, that the human vocal tract is capable of producing only a limited number of different sounds; second, that a speech signal, which is split into small frames, can be re-synthesized by stringing together other frames of audio, if they just sound equally enough. Therefore, the collection of frames of audio used for the re-synthesis has only to be large enough that for every possible audio frame of a speech signal an audio frame that sounds similar enough is available. MFCCs have proven to be a good measure of the perceived similarity of audio content; hence we started from the following basic approach: The code is a string of MFCC vectors. The decoder employs a broad collection of pre-analyzed audio frames. It looks up the audio frame which is closest to the current MFCC vector given in the code. The sequence of these frames is stringed together.

We tested these initial assumptions by generating a first prototype implementation of a system that re-synthesizes a recorded speech signal from a large collection of different speech recordings. As the plain audio waveforms cannot simply be stringed together, as this would result in discontinuities in phase as well as in the waveform, we analyzed and synthesized with half-overlapping windowed frames of 20 ms length. This first test supported our assumption, as we were able to produce intelligible speech output from the system. Here, the intelligibility seemed to be depending on the size and the entropy of the collection of audio frames the re-synthesis was based on. The naturalness of the speech output seemed to improve if recordings of the same speaker were used for the re-synthesis.

3.1. Generating the Lookup Table

The lookup table is generated as follows: During the encoding process, each incoming audio frame is analyzed and the respective MFCC vector is computed. Starting with an empty lookup table, the encoder searches the table for the nearest neighbor MFCC vector. If the lookup table does not contain a vector with a distance smaller than a given threshold, it adds the currently analyzed audio frame and the respective MFCC vector to the lookup table. The threshold for the minimum similarity distance measure is to be chosen by the user in order to determine the coverage in terms of minimum sounding difference of the look-up table.

The look-up table in our prototype is of course fully dependent on the selected speech file. For a real-world codec application, a standardized procedure for generating the look-up table would be sufficient.

3.2. Nearest-Neighbor Search

Each step of the encoding process requires a nearest neighbor search on the MFCC vectors. For that we used

a multi-dimensional search tree [16]. However, basically any other multidimensional nearest-neighbor search technique could be applied at this stage of the algorithm.

3.3. Synthesis Method

To synthesize a continuous-phase output waveform, we decided to use simply half-overlapping windowed frames of 20 ms length. Other voice synthesis methods, as for instance modeling sinusoidal components as described in [14], and also implemented in the analysis / synthesis software SPEAR [17] promise far better results. Such a synthesis method, as it maintains phase accurateness, would also enable analysis and synthesis without overlapping frames, which would make it more applicable in a real-world, low latency codec.

3.4. Size Estimation of the Lookup Table

In the early design stage of this approach, the lookup table is generated new for each audio file to be coded. Hence, for a codec application the lookup table would need to be transmitted to the receiver alongside the indexes. In our early tests, we experimentally tuned the distance measure to gain intelligible results. Using three speech recordings of 45 minute's length of male and female voice characters, we estimate that the percentage of new entries to the lookup table over time can be described with a negative logarithmic function.

In Fig. 1., the average percentage of new entries over time for each 100 analysis steps is plotted, alongside the corresponding logarithmic estimate. After about 20 minutes, only 20% of the analyzed audio frames are added to the lookup table. For the three 45-minute long recordings, the lookup tables' waveform data have approximately 20% the size of the original waveform data, i.e. 80% of the analyzed audio content was not added to the lookup table.

4. EVALUATION

4.1. Bandwidth Estimation

The implementation of the codec idea in the early design stage of the time writing this paper utilizes 32-bit indices for the lookup table of waveform snippets. If only the index for each 20ms frame of analyzed input signal would need to be transferred, i.e. the lookup table is fixed at the sender and receiver, the codec could operate at a bit rate of 1.6 kbit/s. Additional lossless compression measures such as run-length encoding (RLE), Huffman coding, or Lempel-Ziv-Welch (LZW) on the transmitted data itself could help to further reduce entropy and the bit rate.

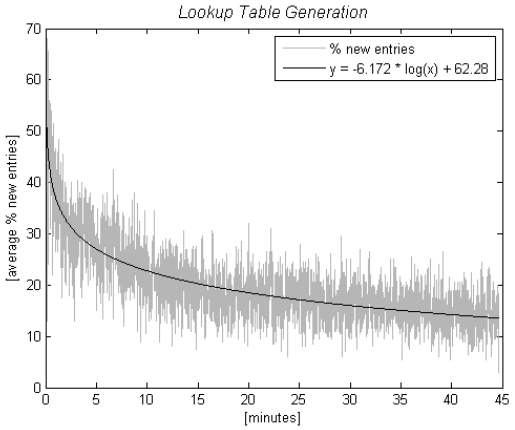

Fig. 1. Average percentage of new entries to the lookup table during the analysis process, and the corresponding logarithmic estimation function.

4.2. Signal Comparison

To demonstrate the abilities of our codec approach, we have re-sysnthesized two random excerpts of each two seconds of a 45-minute long male and female speech recording.

Fig. 2-3. show a comparison of the waveform envelope, the frequency spectrum, and the MFCC feature vectors over time for the original and the re-synthesized signal of each a male and a female human voice recording. In order to present a consistent scale, the first coefficient, which in effect describes the waveform's envelope of the signal, is not plotted, as it is in orders of magnitude larger as the displayed following 25 coefficients.

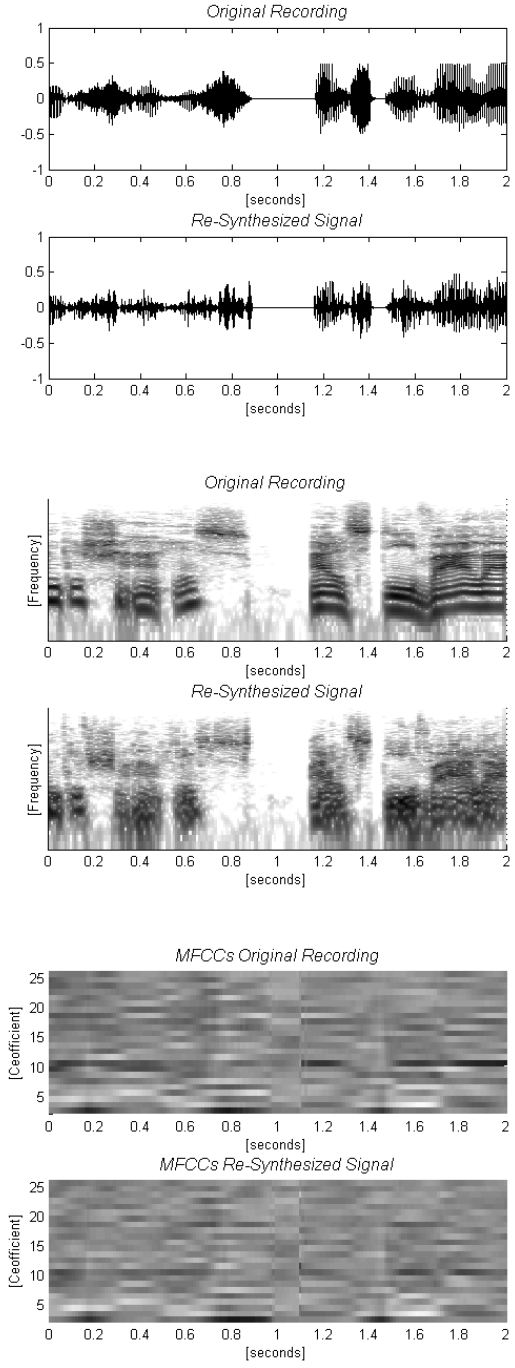

Fig. 2. Comparison of the audio waveform, frequency spectrum and MFCC feature vectors (‘1-25’, ‘0’ not displayed) over time of the original male speech recording and the re-synthesized signal.

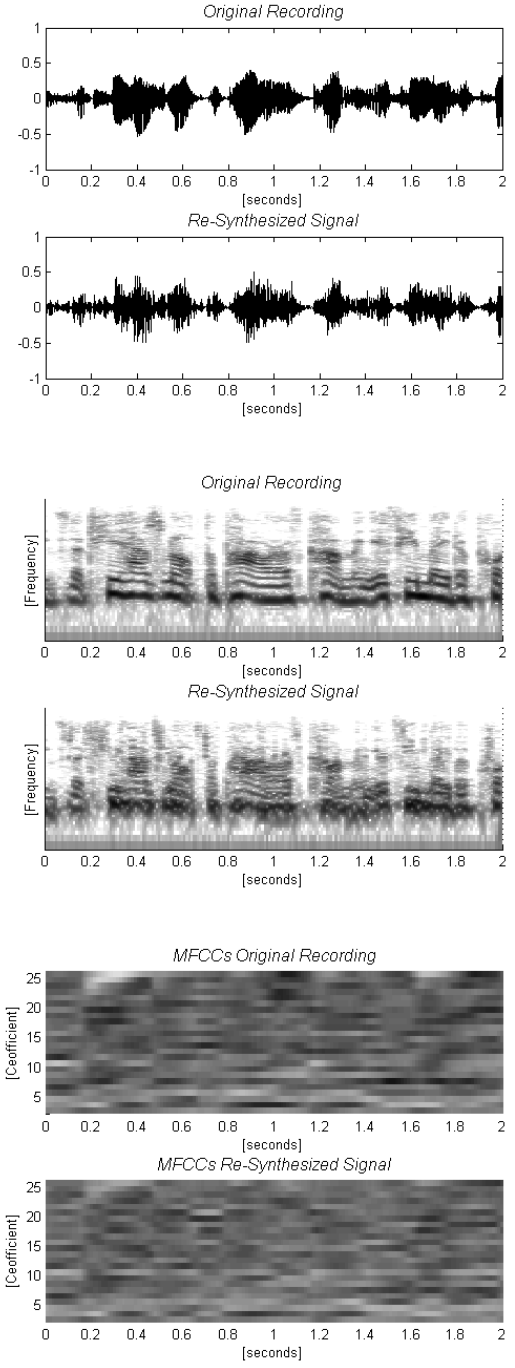

Fig. 3. Comparison of the audio waveform, frequency spectrum and MFCC feature vectors (‘1-25’, ‘0’ not displayed) over time of the original female speech recording and the re-synthesized signal.

5. CONCLUSION AND OUTLOOK

The perceived quality of the synthesized audio output is in our very early version of the software of course dependent on the size and the entropy of the lookup table. Our observations indicate that lookup tables generated from a recording of one speaker results in well-sounding results in the codec for this speaker, lookup tables created from a recording of another speaker however failed to provide appropriate audio content to establish a good-sounding result. Building a unified lookup table is therefore of great concern, especially if it comes to actual use as a codec, where identical lookup table have to be known to both the sender and the receiver. Another possibility would be to use a fixed and an additional adaptive lookup table, similar to the excitation signals of the CELP codec. The fixed lookup table would provide all standard sounds, while the adaptive lookup table would be created during coding, and could especially be tuned to improve the naturalness of the sound of the reproduced speech.

The use of the audio similarity comparison in a real-time codec application as such only became possible by implementing the similarity comparison via the sufficiently fast neighbor search tree algorithm.

Future research should look into other approaches of synthesizing the audio output of the codec, such as modeling sinusoidal components. More work could also be directed to compressing the list of indices to transfer. Standard Run-Length Encoding (RLE) algorithms, as well as statistical models could further reduce the required codec bandwidth.

6. ACKNOWLEDGEMENTS

This work was supported by the “EXIST – University-Based Business Start-Ups” grant 03EGSHB010.

7. REFERENCES

[1] G.711: Pulse code modulation (PCM) of voice frequencies. http://www.itu.int/rec/T-REC-G.711/e. Accessed 2010-09-02.

[2] Hirata, Y. and Nagawa, S. 1989. A 100bit/s speech codec using a speech recognition technique. Proc. EUROSPEECH’89, p. 290-293.

[3] Masuko, T., Tokuda, K. and Kobayashi, T. 1998. A Very Low Bit Rate Speech Coder using HMM with Speaker Adaptation. Proc. ICSLP 1998, pp. 507–510.

[4] Vary, P., Hofmann, R., Hellwig, K., and Sluyter, R. J. 1988. A regular-pulse excited linear predictive codec. Speech Commun. 7, 2 (Jul. 1988), pp. 209–215.

[5] Schroeder, M., Atal, B. 1985. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85. (Apr. 1985), pp. 937–940.

[6] Speex: A Free Codec For Free Speech. http://www.speex.org/. Accessed 2010-09-02.

[7] Codec Quality Comparison. http://www.speex.org/comparison/. Accessed 2010-09-02.

[8] Davies, K.H., Biddulph, R. and Balashek, S. 1952. Automatic Speech Recognition of Spoken Digits, J. Acoust. Soc. Am. 24(6), pp. 637–642.

[9] Mermelstein, P. 1976. Distance measures for speech recognition, psychological and instrumental. Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed,. pp. 374–388. Academic, New York.

[10] Lee, G.H., Yoon, J.S. and Kim, H.K. 2005. A MFCC-based CELP speech coder for server-based speech recognition in network environments. INTERSPEECH-2005, pp. 3169–3172.

[11] Mueller, M. 2007. Information Retrieval for Music and Motion. Springer., p. 65.

[12] Foote, J.T. 1997. Content-Based Retrieval of Music and Audio. In C.-C. J. Kuo et al., editor, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138–147, 1997.

[13] Stevens, K. N., Kasowski, S., and Fant, G. 1953. An electrical analog of the vocal tract. J. Acoust. Soc. Am. 25, pp. 734–742.

[14] McAulay, R. J. and T. F. Quatieri. 1986. Speech Analysis/Synthesis based on a Sinusoidal Representation. IEEE Transactions on Acoustics, Speech and Signal Processing 34(4), pp. 744–754.

[15] Kleinwaechter, T. 2006. Re-synthesis of speech from ASR features. MSc Thesis XR-EE-SIP 2006:005, KTH Stockholm, Electrical Engineering.

[16] Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. Commun. ACM 18, 9 (Sep. 1975), pp. 509–517.

[17] SPEAR homepage. http://www.klingbeil.com/spear/. Accessed 2010-09-02.