HIDDEN MARKOV MODEL BASED AMHARIC SPEECH SYNTHESIZER

Yitagessu Birhanu, Guntram Strecha, Rüdiger Hoffmann

Technische Universität Dresden, Institut für Akustik und Sprachkommunikation
Yitagessu_Birhanu.Gebremedhin@tu-dresden.de

Abstract: This paper explains initial results of a research work on HMM based Amharic synthesis system using phonemes as the fundamental acoustic unit. A speech corpus uttered by a single speaker, and consisting of all diphones of the Amharic language is used for training. It is recorded at 16 kHz sampling rate and 16 bit sample depth, in a standard acoustic studio. Initial results show that the phoneme HMM based Amharic speech synthesizer has acceptable intelligibility and naturalness.

1 Introduction

Amharic is a Semitic language and the national language of Ethiopia. The majority of the 27 million or so speakers of Amharic can be found in Ethiopia, but there are also speakers in a number of other countries like Eritrea, Canada, the USA, and Israel. The name Amharic comes from the district of Amhara in northern Ethiopia, which is thought to be the historic center of the language. Amharic is written with a version of the Ge'ez script known as Fidel. It has 33 basic characters, each of which has seven forms depending on which vowel is to be pronounced in the syllable. There are a number of ways to transliterate Amharic into the Latin alphabet, including but not limited to, one developed by Ernst Hammerschmidt and the EAE Transliteration system. Now days, one can find trainable synthesis systems for Japanese, English, German and some other languages. In the case of Amharic (though it is the official language of a country of more than 80 million people) research work on speech related technologies in general, and text to speech converters in particular are in their infant stages. Few experiments using diphone concatenation synthesis have been reported so far in the field of speech synthesis.
In this research work, we will build the required components from scratch and implement an HMM-based Amharic speech synthesizer by taking into account the particular characteristic of the Amharic language. The main objective is to develop phone and syllable based HMM synthesizers and compare their performances. Unlike other languages, the idea of using a syllable as the basic acoustic unit is feasible as there are only 196 distinctly pronounced Consonant-Vowel (CV) syllables in Amharic. The speech corpus is recorded in a standard acoustic studio at 16 kHz sampling rate. It has a single male speaker and has already been manually labeled. Cepstral features with delta and delta-delta elements are extracted from the speech files and will be used to train the acoustic models. The synthesizer is being developed using the UASR (Unified Approach to Speech Synthesis and Recognition) toolkit of TU Dresden.

2 The HMM based acoustic synthesis

The complete system consists of a training part with the inventory generation and the synthesis part [11]. The parameters of the statistical models used in the system are estimated in the training part. Initially, the speech parameters related to the spectrum are extracted. The spectral parameters consist of the mel-cepstral coefficients and the delta and delta-delta coefficients.

These parameters are then used to train the acoustic models that represent the phonemes. The HMM models are used to generate the diphone inventory. The inventory consists of the optimal state sequences of the diphones measured by a forced Viterbi alignment of the training database. This inventory is used by the synthesis part to build a Gaussian sequence according to the desired input phoneme sequence by concatenating the diphone state sequences of the inventory. We use the MLSA filter of [8] to synthesize the mel-cepstrum sequence taken from the means of the probaility density functions (codebook) of the generated Gaussian sequence.

3 The Database

A speech corpus uttered by a single speaker is used to train the acoustic models. The corpus consists of more than 640 phonetically balanced sentences. It is carefully prepared so that all of the 1156 diphones of Amharic are covered in a fairly equal proportion. The speech signals are recorded in a standard studio at a rate of 16 kHz and quantified to 16 bit.
The speech signals are windowed by a 25.6 ms Hamming window with a 5 ms shift, and then cepstral coefficients were obtained using cepstral-coefficient analysis. The feature vectors consist of 24 cepstral coefficients including the 0^{th} coefficient, and their delta and delta-delta features. Table 1 below shows frequecy of occurence for some of the diphones in the training data.

Diphone	Frequency
ya	23
rA	25
Al	64
ar	34
as	25
da	32

Table 1 - Frequency of occurrence of some diphones in the database.

4 The Training Part and Inventory Generation

For the inventory generation the following steps are necessary (see Figure 1):

1. Perform an HMM training,
2. Determine the optimal HMM state sequences of the carrier words through forced aligment and cut diphone candidates,
3. Choose among duplicate diphone candidates considering the alignment log-likelihood,
4. Encode the chosen diphone inventory:
   - Generate and compress the codebook,
   - Compress the index sequences of the diphones.

The HMM based Amharic acoustic synthesis is developed using phonemes as the basic acoustic units. A 3-state left-to-right HMM topology with 32 Gaussians per state is used to model each phoneme. The speech files are manually labeled in order to have more accurate acoustic models. When the manual labeling process is completed, the HMM training program is provided with

a list of the 34 distinct phonemes, the speech signals, and the label files. It then develops an acoustic model for each of the phonemes.
To build the inventory we determine the sequence of Gaussians with the greatest log-likelihood for each speech file. A speech file is represented by a sequence of feature vectors (mel-cepstrum) and by a corresponding phoneme sequence. We perform a forced Viterbi alignment and automatically cut the resulting HMM state sequence of the speech files into diphone segments. From the alignment procedure we obtain a mean log-likelihood of each diphone segment. If we have duplicate diphones after analyzing all speech files, we select those with the highest mean log-likelihoods. The sequences of Gaussian PDF's representing the selected diphones is the stored to the inventory. We also need to store the codebook containing the means of the Gaussian PDF's into the inventory. The amount of codebook vectors depends on the amount of Gaussians determined by the selected diphones. As each codebook entry is assigned to a particular phoneme we sort the codebook by phonemes and reorder the indices to reduce the bit width needed to store the index sequence.

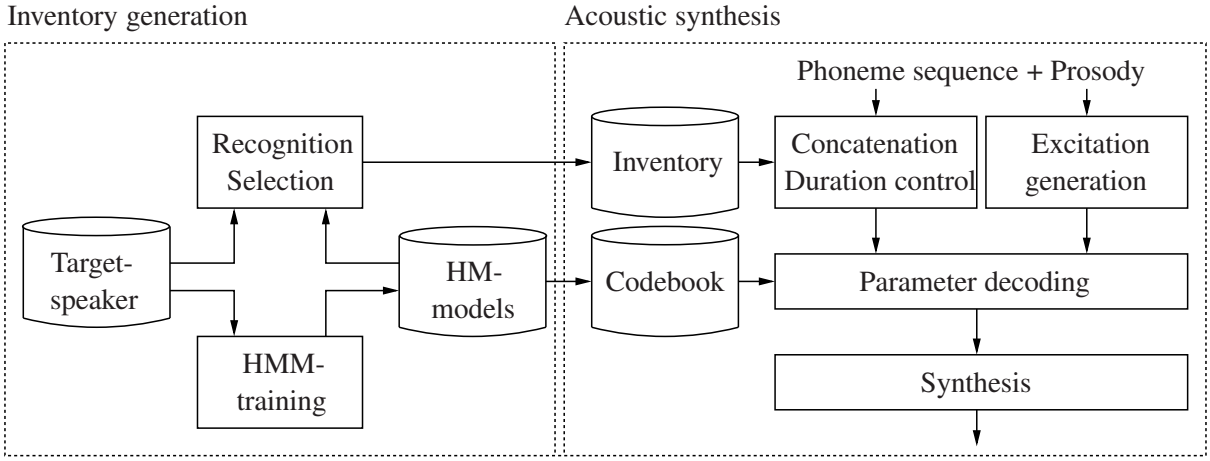

Figure 1 - HMM based synthesis system. Scheme of the inventory generation and the acoustic synthesis.

5 The Synthesis Part

The input to the acoustic synthesis is the desired phoneme sequence annotated with the phone durations and a fundamental frequency contour. The synthesis first concatenates the proper diphone sequence and determines the associated HMM state index sequence. Then the sequence of parameters is generated by a codebook lookup. For scheme A these parameter sequences are the filter parameters of the synthesis filter. We use the MLSA filter [8] together with the perceptual weighting filter proposed in [9]. A block diagram of the acoustic synthesis is shown in Figure 1.

6 Experimental Results, Discussion and Conclusion

In this paper, we explained HMM based Amharic synthesis system using phonemes as the basic acoustic unit. A corpus from a single male speaker is used as the training data. Training of the models is performed using the UASR (Unified Approach to Speech Synthesis and Recognition) Toolkit, which is developed at Technische Universität Dresden, Institut für Akustik und Sprachkommunikation. The output vectors are generated by the HMM based acoustic synthesis

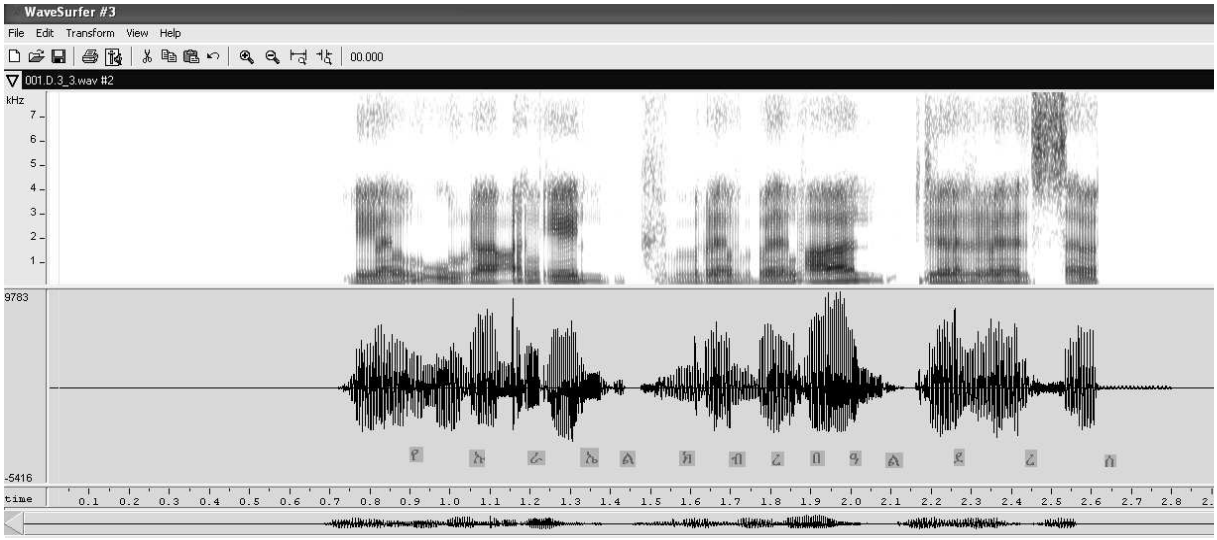

Figure 2 - Spectrogram and waveform of a synthesized speech.

and consists of cepstral coefficients, including the 0^{th} coefficient. To inspect performance of the synthesizer, we have supplied the synthesizer with phoneme sequences of random Amharic texts together with prosodic information taken from a natural representation of these texts. We observed that it generates a synthetic speech with an acceptable quality. Figure 2 shows spectrogram and waveform of an Amharic speech generated by synthesizer.
Initial evaluation results show that phoneme HMM based synthesizer produces acceptable naturalness and intelligibility. It can be concluded that the approach is very effective in rapid development of the acoustic synthesis for new languages. Although quality of the synthesized speech needs to be improved further, its content can be understood without any problem.

7 Future plan – build Syllable HMM based Amharic TTS

Further work on developing Syllable HMM based synthesis and comparing its performance with that of Phone HMM based synthesis will be undertaken. Evaluation of intelligibility and naturalness of the synthetic speech will be done using experts and native Amharic speakers. We also plan to study the trend of the MOS score curve by varying size of the speech corpus.

References

[1] T. Anberbir and T. Takara. Rule based amharic speech synthesis using cepstral method. In *Acoustic Society of Japan (ASJ) Autumn Meeting*, pages 383–384, Sendai, Japan, 2005.

[2] T. Anberbir and T. Takara. Amharic speech synthesis system and its applications to multimedia and telecommunications. In *International workshop on Advanced Image Technology*, pages 186–191, Naha, Okinawa, Japan, 2006.

[3] T. Anberbir and T. Takara. Amharic speech synthesis using cepstral method with stress generation rule. In *Proc. of International Conference on Spoken Language Processing (ICSLP), Interspeech*, pages 1340–1343, Pittsburgh, 2006.

[4] T. Anberbir and T. Takara. Amharic spoken word synthesis using stress generation rule. In *Acoustic society of Japan (ASJ) 2006 Spring Meeting*, pages 329–330, Tokyo, 2006.

[5] T. Anberbir and T. Takara. Development of an amharic text-to-speech system using cepstral method. In *Proc. of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT)*, page 46–52, Athens, 2009.

[6] T. Anberbirand, M. Gasserand, T. Takaraand, and K. D. Yoon. Grapheme-to-phoneme conversion for amharic text-to-speech system. In *Conference on Human Language Technology for Development (HLTD)*, pages 68–73, Alexandria, 2011.

[7] T. Anberbirand, T. Takara, and D. Y. Kim. Modeling of geminate duration in an amharic text-to-speech synthesis system. In *Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 122–129, Penang, Malaysia, 2010.

[8] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, volume 8, pages 93–96, Apr. 1983.

[9] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai. Efficient encoding of mel-generalized cepstrum for celp coders. *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, 2:1355–1358, Apr. 1997.

[10] S. H. Mariam, S. P. Kishore, A. W. Black, R. Kumar, and R. Sangal. Unit selection voice for amharic using festvox. In *5th ISCA Speech Synthesis Workshop*, pages 103–108, Pittsburgh, 2004.

[11] G. Strecha and M. Wolff. Speech synthesis using hmm based diphone inventory encoding for low-resource devices. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5380–5383, Prague, 2011.

[12] N. Tademe. *Formant Based Speech Synthesis: Synthesizing Amharic vowels*. VDM Verlag, Saarbrücken, 2009.

[13] T. Yoshimuraand, K. Tokudaand, T. Masukoand, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Proc. of Eurospeech*, page 2347–2350, 1999.