Design and First Implementation of a Spoken Language Interface for Romanian Language

ALINA NICA ALEXANDRU CARUNTU GAVRIL TODEREAN
Department of Telecommunications
Technical University of Cluj-Napoca
26-28 Baritiu str., Cluj-Napoca
ROMANIA

Abstract: - Spoken language interfaces offer a great potential for enhancing human-computer interaction, speech being the most natural and efficient manner to exchange information for most of us. Such systems consist in automatic speech recognition and text-to-speech synthesis engines, together with a language modeling module. All of these parts of the system require a great amount of computing time and memory space, so manners of using some parts of a module to build another must be investigated in order to optimize the functionality of the whole system. This paper presents the architecture of a spoken language interface, and some implementation details and practical results obtained both for speech recognition and text-to-speech synthesis.

Key-Words: - Spoken language interface, automatic speech recognition, text-to-speech synthesis, language modeling, hidden Markov models, mel-frequency cepstral coefficients.

1 Introduction

Although there are a large number of modalities by which a person can interact with a machine (text, graphical, touch screen, mouse, etc.) it is obvious that the most natural one of them is speech. *Spoken language interfaces* are required not only by impaired people, but also by regular users, who want an easier access to their computers [1], [2].

Over the last decades significant advances have been made in the fields of automatic speech recognition (ASR), text-to-speech synthesis (TTS) and language technologies. More and more conversational systems are leaving the laboratories to take their place in customer service centers or telephone-based information systems. Despite all of these we are far from completely solving all the problems in this area and far from the performances of the humans. Currently there is no spoken language system that is able to fully recognize conversational speech, or to read as natural as a human speaker.

As Figure 1 shows, a spoken language interface consists mainly of three parts: automatic speech recognition, text-to-speech synthesis, and language modeling, so we are dealing with a very complex system. This means that we need lots of input data and a significant amount of time for training, both for recognition and synthesis. Also, each of these modules taken separately requires significant memory resources for data storage, and putting them together would only make things worse in terms of memory needs. It results from here that we must find a way to reuse some parts of a module to another in order to optimize the whole system. This is not an easy job since the principles that govern these two processes are different in terms of technology.

But from which module to "borrow" parts which can be used by the other one, in order to improve the functionality of our spoken language interface? Many studies have been conducted in recent years aiming to find an answer to this question. All that we can say so far is that the trend is to use mostly speech recognition technologies in order to synthesize speech, although attempts have been made from the other direction too [3], [4]. Only one thing seems to be certain no matter the approach: the hidden Markov models (HMM) are the heart of such unified system. Already a state-of-the-art technology in the speech recognition field, they have been used intensively in the last decade for speech synthesis also [5]–[8].

A last problem that needs to be solved is that of parameters used for the representation of speech signals. We decided to use the mel-frequency cepstral coefficients (MFCC). These are the features that give the best results in the case of speech recognition, and have been tested successfully in TTS systems also [8], [9].

This paper summarizes our recent work in the areas of speech recognition and speech synthesis in an attempt to build a spoken language interface. We present mostly a high-level perspective of our research, details of particular aspects being given when needed.

The speech recognition engine has been developed in C++ and recognition rates for isolated words in Romanian are

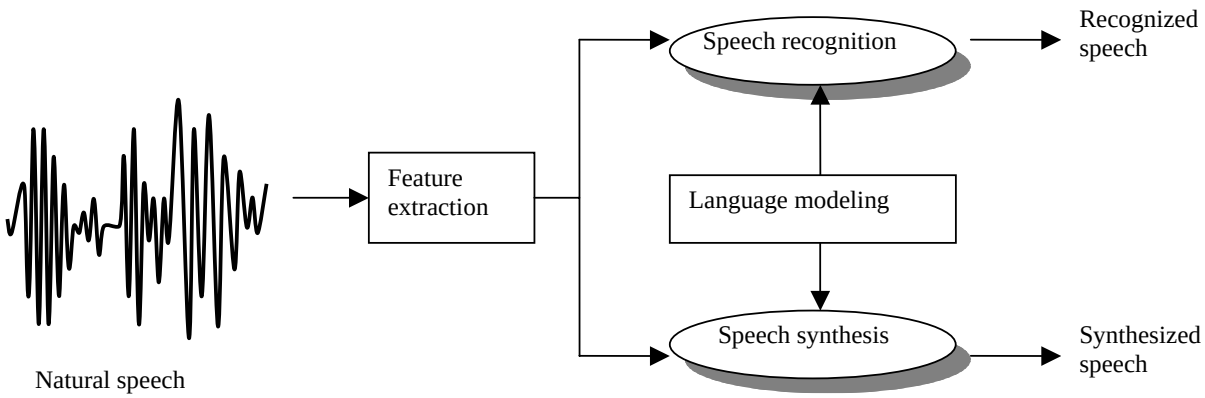

Figure 1. The block-diagram of a spoken language interface

presented. In the case of synthesis we describe our first experiments using MATLAB, regarding the use of MFC coefficients for synthesis. We will present here a solution of reconstructing the speech signal using MFCC, which we intend to use in our spoken language interface.

2 Spoken Language Interface Modules

As already mentioned in the introductory part of this paper, a spoken language interface consists of three main parts: automatic speech recognition, speech synthesis and language modeling. We will present here only the first two of them since no research has been carried out yet by our team for language modeling. This is mainly due to the fact that currently there is no standard speech database for the Romanian language. This is also the reason why we tested our system on isolated words instead on continuous speech. But first, a few words about mel-frequency cepstral coefficients, used for signal representation.

2.1 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients are the most used features for speech recognition. As pointed out in [10], there are several ways of computing the MFCC:

- The solution proposed in 1980 by Davis and Mermelstein [11]: a filter bank of 20 filters, a sampling frequency of 10 kHz, and a speech bandwidth between 0 and 4600 Hz.
- The implementation from HTK, described in [12]: 24 filters, sampling rate greater than 16 kHz, [0, 8000] Hz speech bandwidth.
- The implementation from Malcom Slaney's Auditory Toolbox for MATLAB [13], which assumes 40 filters in the filter bank, a sampling rate of 16 kHz, and a speech bandwidth between 133 and 6854 Hz.

We have chosen to implement the last one in C++ for speech recognition and we used the MATLAB version (which is freely available at location [14]) for speech synthesis. This way, the compatibility between our experiments is assured.

In Slaney's version the centre frequencies of the filter bank are spaced as follows [13]:

- The centre frequencies of the first 13 filters are linearly spaced in the range [200, 1000] Hz at 66.67 Hz one from the other.
- The centre frequencies of the last 27 filters are logarithmically spaced between 1071 and 6400 Hz, with a step of 1.0711703 Hz.

Also, the filter bank is normalized in such a way that the sum of coefficients for every filter equals one.

2.2 Speech recognition

The speech recognition engine is implemented in C++ and is based on hidden Markov models. Most of the speech recognizers developed nowadays is based on HMM technology, because of their capability of best modeling the statistical nature of the speech signals. Shortly, a HMM is described by three parameters: the initial state distribution, the state transition probability distribution, and the observation symbol probability distribution (or output probabilities) in a certain state. Each of them is represented by a matrix, which must be estimated. A complete reference of the algorithms used for this task can be found in [15].

We used for our experiments continuous HMMs, which means that they are using mixtures of Gaussians to characterize the model transitions, so we had to estimate another three parameters in this case: the weights of the gaussians in the mixture, the means and the covariances. For the experiments presented in this paper we used diagonal

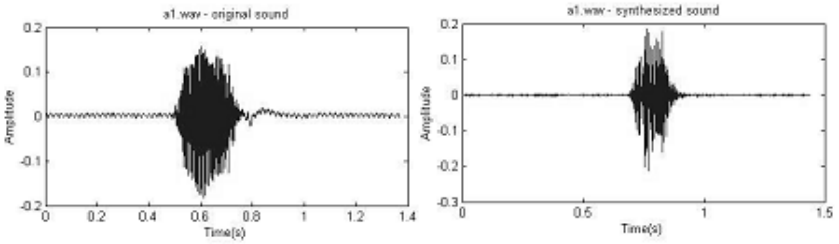

Figure 2. The original and synthesized sound for the vowel *a* (a1.wav)

covariance matrices, since they are defined with far less parameters, so we need less acoustic data to train them. Also, they are easier to estimate than full matrices.

As pointed out before, no language constraints have been applied since the experiments were carried out on isolated words.

2.3 Speech synthesis

Nowadays, most of the existing text-to-speech systems can synthesize good quality speech, but the problem is that they cannot synthesize speech with various voice characteristics (speaking styles, emotions, etc).

For the speech synthesis systems which are based on the selection and concatenation of acoustical units, a large amount of speech databases is required. In the case of the TTS systems where speech can be generated directly from hidden Markov models (HMM) the voice characteristics modifications can be done without being necessary large databases, by applying, e.g., a speaker interpolation technique or a speaker adaptation technique. The HMM-based speech synthesis system (HTS) was implemented for the first time for Japanese language; it has also been applied to other languages, e.g., English, Portuguese [16], [17].

In order to obtain different speaking styles such as emotion expression, pitch and duration parameters have to be accordingly modified. A HMM-based speech synthesis system models not only the spectrum parameter but also the fundamental frequency and duration in the unified framework of HMM [16]. As spectral parameters are used mel-cepstral coefficients, which are obtained from a speech database using a mel-cepstral analysis technique. The speech waveform can be synthesized directly from the generated mel-cepstral coefficients and fundamental frequency by using the MLSA (Mel Log Spectrum Approximation) filter [8].

The mel-cepstral analysis technique is a kind of cepstrum with a Mel scale. The Mel scale takes into account the frequency resolution properties of the human ear. The MFCC are obtain as follows: after preemphasis and windowing, the spectrum of the speech signal is computed and then translated to mel scale with the help of the filter bank described in paragraph 2.1. The final step is the discrete cosine transform (DCT) applied to the logarithms of mel frequencies, and we obtain the MFC coefficients.

In order to synthesize sounds this process must be inverted. This is not an easy job since during the process a lot of informations are lost. For example, the phase information is discarded when taking the FFT spectrum. Also, usually, in the final stage, a number of coefficients are truncated, not saying that the mel-filterbank quantises the frequency spectrum into a reduced resolution [18].

Despite all of these it is possible to recover a reasonable estimate of the magnitude spectrum, in the following manner. First of all the truncated MFC coefficients are padded with zeroes to the length of the filter bank (in this work the length of the filter bank is 40 and the number of MFCCs is 13). An inverse DCT is applied to MFC coefficients, followed by an exponential scaling (the inverse operation of the logarithm), and we obtain an estimate of the mel-filter bank vector. The final stage is the inverse translation from the mel-frequency domain to spectral domain.

3 Experiments and results

The experiments that we have conducted can be divided into two classes: test involving speech recognition and test involving speech synthesis. The speech was parameterized using a preemphasis filter with a coefficient of 0.97, with a Hamming window of 25 ms length and a step of 10 ms. The MFCC were extracted using a 40 channels filter bank and only the first 13 coefficients were used. The recorded speech signal was sampled at a rate of 16 kHz.

For speech recognition we used a database consisting in utterances of the digits from 0 to 9 in Romanian language. A number of tests have been performed with different values for the parameters of the HMMs and the results ranged between 86% and 94% recognition rates. For the task of testing our HMM library these values can be considered satisfactory.

For speech synthesis we used the Romanian vowels uttered by several speakers (males and females) and we reconstructed them using the MFCCs. The result of this operation, for the vowel *a* can be seen in Figure 2. In the left side we have the original signal, while in the rigth side there is the waveform of the signal obtained from the reconstruction of the spectral shape from the MFCCs. It can be observed that the two signals have similar shapes but the

synthesized signal, although intelligible, sounds like whispering because we did not use the pitch information at this stage.

4 Conclusions

A number of experiments involving a spoken language interface for Romanian language were presented here. This kind of system is based on two different technologies from speech processing field, namely recognition and synthesis. Putting them together generates a number of problems that needs to be solved since those technologies have few things in common.

First of all the tool on which those processes are based are the HMMs. Because they are known as a speech recognition technology we performed a number of tests (isolated words recognition) in order to evaluate a libray that we have implemented in C++. The recognition rates that we have obtained can be considered satisfactory which means that we can use it in further tests involving the synthesis.

A second problem that arises when trying to merge synthesis with recognition is that of the features used to represent the speech signal. We started with the state-of-the-art parameters used in speech recognition (MFCCs) and we tried to invert them and use to reconstruct the speech signal. The sounds that we have obtained in this manner sounds like whispering because we did not incorporate the pitch estimates. For a standard speech recognizer this information is not important but in order to obtain a high-quality resynthesis this is strongly needed.

In the near future we intend to improve the quality of our synthesizer and to transform our isolated words recognizer into a continuous one.

References:

[1] Dusan, S, Flanagan, J, A System for Multimodal Dialogue and Language Aquisition, *Proceedings of the 2nd Conference Speech Technology and Human-Computer Dialogue*, Bucharest, April 10-11, 2003, pp. 51-60.

[2] Burileanu, D, Fecioru, A, Ion, D, On Automatic Speech Synthesis for Spoken Language Interfaces, *Proceedings of the 2nd Conference Speech Technology and Human-Computer Dialogue*, Bucharest, April 10-11, 2003, pp. 127-138.

[3] Chen, K, Borys, S, Hasegawa-Johnson, M, Prosody Dependent Speech Recognition with Explicit Duration Modelling at Intonational Phrase Boundaries, *Proc. EUROSPEECH*, 2003, pp. 393-396.

[4] Shriberg, E, Stolcke, A, Prosody Modeling for Automatic Speech Recognition and Understanding, *Mathematical Foundations of Speech and Language Modeling, M. Johnson, M. Ostendorf, S. Khudanpur, R. Rosenfeld (eds.)*, Volume 138 in IMA Volumes in Mathematics and its Applications, pp. 105-114, Springer-Verlag.

[5] Falaschi, A, Giustiniani, M, Verola, M, A Hidden Markov Model Approach to Speech Synthesis, *Proc. EUROSPEECH*, 1989, pp. 187-190.

[6] Giustiniani, M, Pierucci, P., Phonetic Ergodic HMM for Speech Synthesis, *Proc. EUROSPEECH*, 1991, pp. 349-352.

[7] Masuko, T, Tokuda, K, Kobayashi, T, Imai, S, Speech Synthesis from HMMs Using Dynamic Features, *Proc. ICASSP*, 1996, pp. 389-392.

[8] Yoshimura, T, Tokuda, K, Masuko, T, Kobayashi, T, Kitamura, T, Simultaneous Modeling of Spectrum, Pitch, and Duration in HMM-based Speech Synthesis, *Proc. EUROSPEECH*, 1999, pp. 2347-2350.

[9] Demuynk, K, Garcia, O, Van Compernolle, D, Synthesizing Speech from Speech Recognition Parameters, *Proc. International Conference on Spoken Language Processing*, volume II, pages 945--948, Jeju Island, Korea, October 2004.

[10] Gancev, T, Fakotakis, N, Kokkinakis, G, Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, *Proc. of the 10th International Conference on Speech and Computer, SPECOM*, 2005, Vol. 1, pp. 191-194.

[11] Davis, S. B, Mermelstein, P, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, *IEEE Trans. on Acoustic, Speech and SignalProcessing*, Vol. 28, No. 4, 1980, pp. 357–366.

[12] Young, S. J, Odell, J, Ollason, D, Valtchev, V, Woodland, P, *The HTK Book. Version 2.1*, Department of Engineering, Cambridge University, UK, 1995.

[13] Slaney, M, *Auditory Toolbox. Version 2*, Technical Report #1998-010, Interval Research Corporation, 1998.

[14] *** Auditory Toolbox, http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/

[15] Rabiner, L. R, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proc. IEEE*, Vol. 77, No. 2, 1989, pp. 257--286.

[16] Tokuda, K, Zen, H, Black, A. W, An HMM-Based Speech Synthesis System Applied to English, *IEEE Speech Synthesis Workshop*, Santa Monica, California, Sept. 11-13, 2002.

[17] Maia, R. da S, Zen, H, Tokuda, K, Kitamura, T, Resende Jr, F. G. V, Towards the Development of a Brazilian Portuguese Text-to-Speech System Based on HMM, *Proceedings of European Conference on Speech Communication and Technology*, Geneva, Switzerland, Sept. 1-4, 2003.

[18] Milner, B. P, Shao, X., Speech Reconstruction from MFCCs Using a Source-Filter Model, *Proc. ICSLP*, 2002, pp. 2421-2424.