2010 IEEE Fourth International Conference on Semantic Computing

Automatic Speech Recognition System Based on Wavelet Analysis

Mariusz Ziółko, Jakub Gałka, Bartosz Ziółko, Tomasz Jadczyk, Dawid Skurzok, Jan Wicijowski
Department of Electronics, AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
www.dsp.agh.edu.pl, {ziolko, jgalka, bziolko}@agh.edu.pl

***Abstract*—We demonstrate an automatic speech recognition system for Polish continuous speech. As most of the progress in the field is done for English, a few layers of our system are different from popular approaches in this field. These elements of our system could be successfully ported to other languages which share some features with Polish: the speech contains a lot of high-frequency phones (fricatives and plosives) and is highly inflective and non-positional.**

I. Introduction

Research on automatic speech recognition (ASR) started several decades ago. It has resulted in many successful designs, however, ASR systems are always below the level of human speech recognition capability, even for English. In case of less popular languages, like Polish (with around 60 million speakers), the situation is much worse. There is no large vocabulary ASR software for Polish. There are some commercial call centre applications (PrimeSpeech) but they are limited in their domain areas and not described in scientific papers. Our system is advanced and targeted for Polish, while others [1], [7], [8] are more general, and strongly based on HTK framework [10].

The following sections describe various levels of our system. The second section outlines phoneme segmentation [3]. The third one depicts perceptual discrete wavelet transform (DWT) applied for parameterisation. The fourth section describes our approach to building sentences from word hypotheses with language models, including semantic one.

II. Speech segmentation

Six levels dyadic decomposition procedure with discrete Meyer wavelet decomposition filters were applied to speech signal to obtain a discrete wavelet power spectrum. The time discretisation for all wavelet sub-bands is unified by summing adequate number of wavelet samples.

The result has to be smoothed with running mean as a low-pass FIR filter with a length of 20 milliseconds. This value is related to assumed length of the shortest speech segment [5].

The segmentation algorithm is based on detection of energy transitions between different wavelet sub-bands. The transitions significant enough are called the events. An event occurs when the energy transition rearranges the order of sorted power bands of the wavelet spectrum.

Segment borders locations are extracted with an event detection function by choosing its local maxima, which fulfill

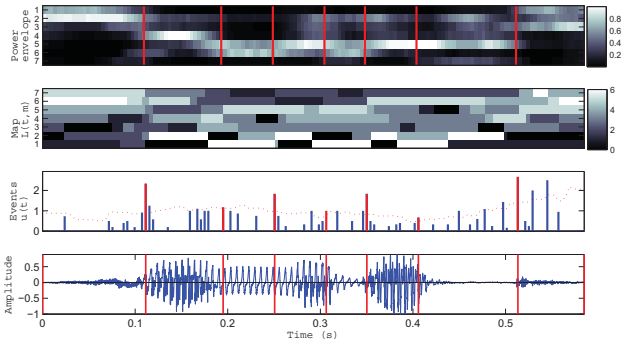

Fig. 1. Exemplary segmentation of the Polish word ”Henryk”

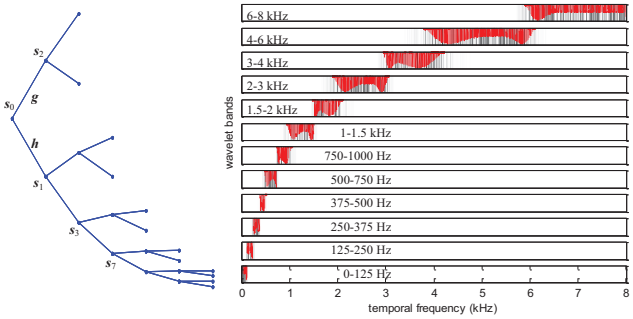

Fig. 2. Perceptual speech feature extraction analysis, based on wavelet decomposition (left). Temporal frequency-sweep response of the decomposition (right)

two basic conditions. Each of the chosen peaks has to be the highest one within its neighborhood, and higher than specified threshold. More comprehensive description of the method was given in [3].

III. Speech parameterisation based on discrete wavelet transforms

In the very most of ASR solutions, filter banks are used for parametrisation of speech into acoustic features. In our system, the wavelet decomposition is applied to improve time efficiency. The time efficiency is a very important factor in large vocabulary ASR systems, because it is difficult to construct an efficient system which is able to analyse enough data and compare it with a whole dictionary in real-time.

978-0-7695-4154-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICSC.2010.16

We have resorted to multithreaded, parallel solutions for the implementation of the software.

Optimal wavelet tree (see Fig. 2) was found to choose exact boundaries of frequency subbands [4]. This approach provides decent psycho-acoustic model of the human ear Mel-like frequency characteristics [2]. The signal s_0 is decomposed with decomposition discrete Meyer filters g and h according to the designed perceptual tree. The parametrisation is conducted by measuring different subband energy fractions and storing them in a vector of their magnitudes.

IV. Level Building and Language Modeling

There are two typical structures which can be used to model a sentence from word hypotheses: N-best list and a word lattice. N-best list is easier in implementation, while a lattice outperforms it thanks to ability of choosing more similar word combinations. We have decided to use a lattice due to the large-scale requirement.

The acoustic classifier provides a stream of phonetic hypotheses to the word decoder level. Word decoder searches for words to match the hypothesis sequences of different lengths, approximately equal to time necessary to pronounce a word. All phonetic hypotheses are evaluated by a word-level decoder by comparing them to words from a dictionary using modified Levenshtein distance (in a pronunciation, phonetic domain) [6]. The most likely version matching a particular word (with respect to its length) is chosen. Next, the algorithm proceeds for a next phonetic hypotheses sequences exactly after the end of the previous chosen word hypothesis. The classifier always yields several parallel word hypotheses with top likelihoods. The algorithm connects them as paths/nodes of the graph, if their beginnings and endings are the same. Then the lattice is ready for further processing.

A typical strategy to search for a best path through a word lattice is by applying Viterbi algorithm [9]. In our case, we first want to reduce the number of edges in the lattice, by pruning the connections between words which do not appear in 2-grams tuples database. The 2-grams are word statistics, collected by us from over 10 GiB of text. They are representative enough so that in most cases it can be assumed, that if there is no 2-gram in the corpus, then two words are prohibited to appear one after another in a correct sentence. This strategy will allow us to reduce a lattice substantially, allowing to conduct calculation in real time, even with a large vocabulary. Due to inflections and non-positionality it is believed that it is not possible to build an efficient 3-gram model for Slavic languages with existing corpora resources. The reason for this is that there are fewer language resources and much more posible word combinations.

Finally, a semantic model is used to reorder the sentence hypotheses. The reordering is presented in a way, which demonstrates real impact of the semantic model on the results. The semantic model is based on bag-of-words scheme by applying a word-topic matrix, which contains statistical data of appearing particular words in topics. A topic can be a text unit ranging from a sentence through a paragraph to a document.

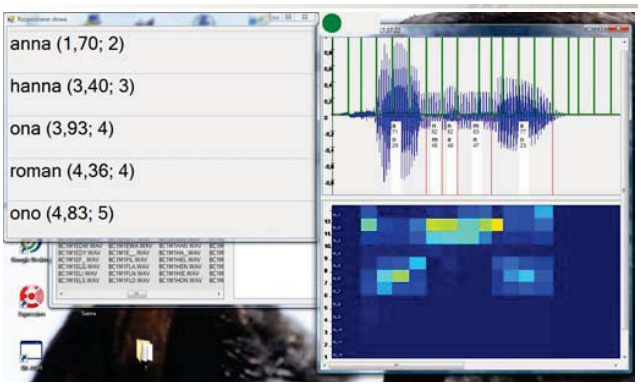

Fig. 3. Screenshot of our automatic speech recognition software

The words can be lemmatised before the matrix creation. All the possible scenarios can be tested independently or as a combination of a few models.

V. Conclusions

Our ASR system (see Fig. 3) could be demonstrated as a solution which is based on different methods then Hidden Markov Model and Viterbi algorithm. The working software was made for demonstrations to present new solutions as well as for development and tests of our algorithms. It shows not only results but also the process of taking particular decisions on different levels presented above.

VI. Acknowledgments

This work was supported by MNISW grant OR00001905.

References

[1] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziski, and N. Cylwik, "JURISDIC Polish speech database for taking dictation of legal texts," *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1280–1287, 2008.

[2] O. Farooq and S. Datta, "Wavelet based robust subband features for phoneme recognition," *IEE Proceedings: Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 187–193, 2004.

[3] J. Gałka and M. Ziółko, "Wavelets in speech segmentation," *Proceedings of The 14th IEEE Mediterranean Electrotechnical Conference MELECON 2008, Ajaccio*, 2008.

[4] ——, "Wavelet parametrization for speech recognition," *Proceedings of an ISCA tutorial and research workshop on non-linear speech processing NOLISP 2009, VIC*, 2009.

[5] S. Grocholewski, "First database for spoken polish," *Proceedings of International Conference on Language Resources and Evaluation, Grenada*, pp. 1059–1062, 1998.

[6] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, pp. 707–10, 1966.

[7] L. Pawlaczyk and P. Bosky, "Skrybot - a system for automatic speech recognition of Polish language," *Advances in Soft Computing, Man-Machine Interactions, Springer*, vol. 59/2009, pp. 381–387, 2009.

[8] A. Pułka and P. Kłosowski, "Polish semantic speech recognition expert system supporting electronic design system," *Proceedings of Conference on Human System Interactions (HSI), Krakow*, vol. 479-484, 2008.

[9] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, April 1967.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK Book*. UK: Cambridge University Engineering Department, 2005.