Automatic Speech Recognition System Dedicated for Polish

Mariusz Ziółko, Jakub Gałka, Bartosz Ziółko, Tomasz Jadczyk, Dawid Skurzok, Mariusz Mąsior

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

An automatic speech recognition system for Polish is demonstrated. A few layers of our system are different from popular approaches as a result of differences between Polish and English languages.

1. Introduction

Research on automatic speech recognition (ASR) started several decades ago. Most of the progress in the field was done for English. 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 (LVR) software for Polish. Polish speech contains very high-frequency phones (fricatives and plosives) and the language is highly inflected and non-positional. There are some commercial call centre applications, developed by PrimeSpeech, but they are limited to their domain areas. Our system is based on modified kNN classifier and wavelets. It is targeted for Polish, while others [1, 6, 7, 5] are more general, and strongly based on HTK framework [9].

2. Speech parametrisation based on discrete wavelet transforms

Eleven levels perceptual decomposition procedure with discrete Meyer wavelet decomposition (WD) filters were applied to obtain a power spectrum of speech signal. The time discretisation for all wavelet sub-bands is unified by summing adequate number of wavelet spectrum samples for high frequencies.

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 an assumed length of the shortest speech segment [4].

The WD is applied, instead of filter banks, to improve time efficiency. It is difficult to construct a system which is able to analyse all data and compare it with a whole dictionary in real-time. A multithreaded solution was implemented.

Optimal wavelet tree (Fig. 1) was found to choose exact boundaries of frequency subbands [3]. The signal s_0 is decomposed with discrete Meyer high-pass filters g and low-pass h according to the designed perceptual tree. This approach provides decent psycho-acoustic model of the human ear Mel-like frequency characteristics [2]. The parametrisation is conducted by measuring different subband energy fractions and storing them in a vector of their magnitudes. The system was trained on CORPORA [4] and GlobalPhone [8].

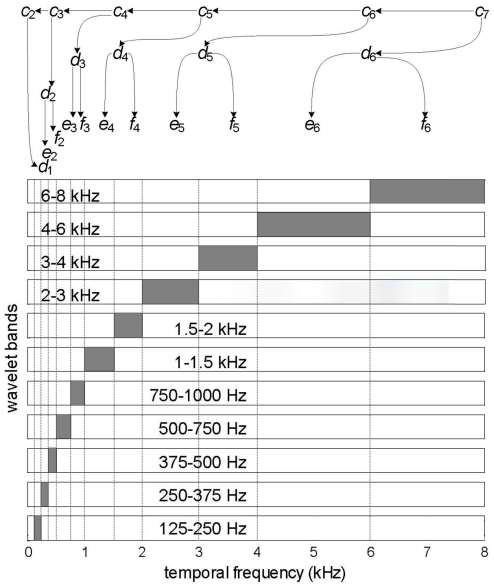

Figure 1: *Perceptual speech feature extraction analysis, based on wavelet decomposition with temporal frequency-sweep response of the decomposition*

3. Level Building and Language Modelling

Parameter vectors are classified using modified k-NN algorithm. The acoustic classifier provides a stream of phonetic hypotheses to a word decoder level.

The word decoder seeks for words to match phoneme hypothesis sequences of different lengths, approximately equal to time necessary to pronounce a particular word. All phonetic hypotheses are evaluated by comparing them to dictionary using modified Levenshtein distance in phonetic domain [10]. The most likely one matching a length of a word is chosen. Then, the algorithm proceeds for a next phonetic hypotheses sequences. The classifier always yields several parallel word hypotheses with top likelihoods. The algorithm connects them if their beginnings and endings are close to each other.

Before searching for a best path through a word lattice our system attempts to reduce a number of edges in a lattice, by pruning connections between words which do not appear in 2-gram model (Fig. 3), collected from over 10 GB 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 excluded 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.

28 – 31 August 2011, Florence, Italy

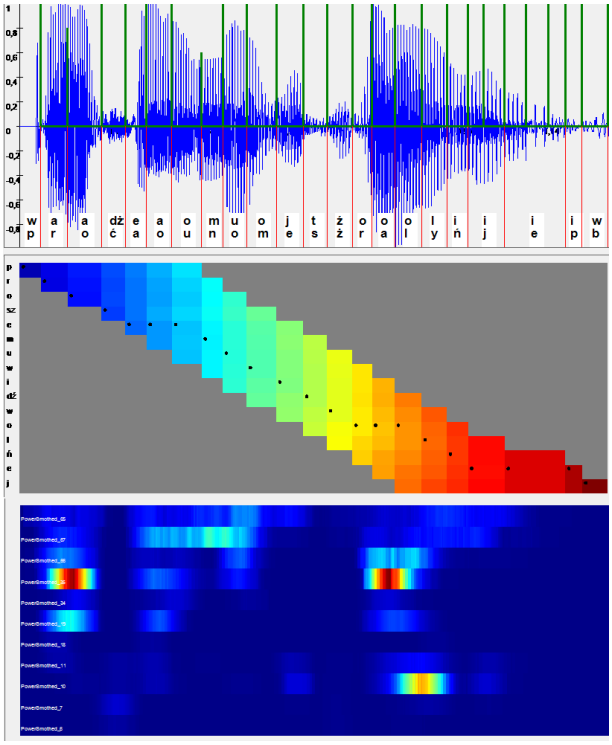

Figure 2: *Signal with phoneme hypotheses (upper), dynamic time warping to fit observed segments to phonemes (middle), energy in frequency subbands (lower)*

4. Conclusions

Our ASR system is a solution based on methods different then HMM. The effective software was made for demonstrations to present new solutions, as well as to develop and to test new algorithms. It shows not only results, but also the process of taking particular decisions on different levels presented above.

Acknowledgements

This work was supported by MNISW grant OR00001905.

5. References

[1] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziński, 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," *IEEE Proceedings: Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 187–193, 2004.

[3] J. Gałka and M. Ziółko, "Wavelet parametrization for speech recognition," *Proceedings of an ISCA tutorial and research workshop on non-linear speech processing NOLISP 2009, VIC*, 2009.

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

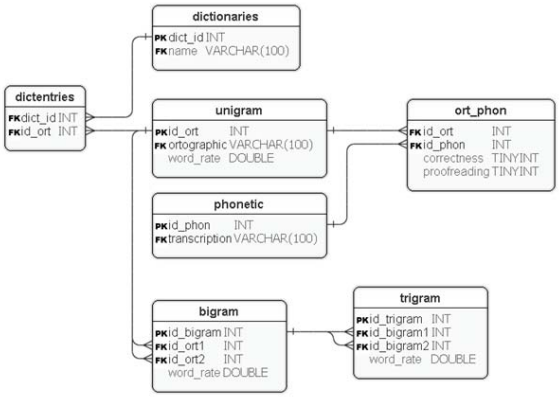

Figure 3: *Dictionary and n-gram model implemented in SQL*

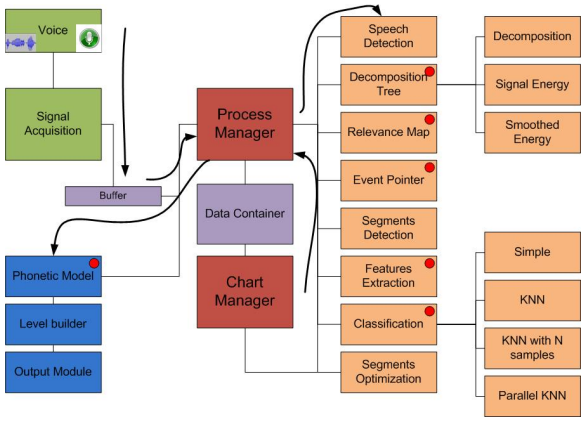

Figure 4: *Scheme of the implementation of our ASR system*

[5] K. Marasek, Ł. Brocki, D. Koržinek, K. Szklanny, and R. Gubrynowicz, "User-centered design for a voice portal," *Aspects of Natural Language Processing, Lecture Notes in Computer Science 5070*, pp. 273–293, 2009.

[6] 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.

[7] 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.

[8] N. T. Vu, F. Kraus, and T. Schultz, "Multilingual astabil: a new confidence score for multilingual unsupervised training," *Proceedings of IEEE Workshop on Spoken Language Technology, SLT 2010, Berkley*, 2010.

[9] 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.

[10] B. Ziółko, J. Gałka, D. Skurzok, and T. Jadczyk, "Modified weighted Levenshtein distance in automatic speech recognition," *Proceedings of XVI KKZMBM*, pp. 116–120, 2010.