Evaluation of Errors in Polish Phones Segmentation for Different Types of Transitions

Bartosz Ziółko, Mariusz Ziółko
Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
www.dsp.agh.edu.pl
{bziolko,ziolko}@agh.edu.pl

Suresh Manandhar, Richard C. Wilson
Department of Computer Science
University of York
Heslington, York, UK
www.cs.york.ac.uk
{suresh,wilson}@cs.york.ac.uk

***Abstract*— The paper presents an evaluation of Polish phone segmentation for different types of phones. The categorisation was done based on acoustic properties. The segmentation method is based on discrete wavelet transform and was already published. The results show that several types of transitions, especially from and to vowels cause more errors than others.**

I. Introduction

Speech signals typically need to be divided into small segments before starting a recognition procedure. Analysis of these frames can determine the likelihood of a particular phone being present within the frame. Speech is non-stationary in the sense that frequency components change continuously over time, but it is generally assumed to be a stationary process within a single frame. Naturally, this causes recognition difficulties if the frame contains the end of one phone and the beginning of another. Segmentation methods currently used in speech recognition do not consider where phones begin and end. Uniform segmentation causes conflicting information to appear at the boundaries of phones. For more accurate modelling, non-uniform phone segmentation can be useful in speech recognition [1].

Even though, there are hand-labelled speech corpora, the automatic speech segmentation has several applications. For example, it can be used during recognition process to model audio units corresponding to whole phones rather than artificial units like 23 ms frames. In Polish phone segmentation is especially important because the phonetic rules are more regular than in English and there are some non-existing in English phones which have very specific frequency content.

Errors in phone segmentation depend on what type of transitions are being detected. The evaluations differ regarding to groups of phones because some phones have similar spectra, while others differ a lot. These differences depend on acoustic properties of phones [2].

Segmentation is a common but challenging task in audio processing. There are many different approaches to this task [3], [4], [5], [6], [7], [8], [9]. These approaches must be evaluated and compared, which is not a very easy task itself due to ambiguity of segmentation. First, it is not a process which can be clearly compared with the ground truth, because it is difficult to do virtually correct segmentation, even manually or with supervision. Additionally the comparison of segmentation is typically not binary. We can rarely say that segmentation is definitely correct or wrong. Usually we can say that it is rather accurate, quite good or bad. We applied the recall and precision evaluation method [10], which is commonly used in information retrieval, using fuzzy sets [11]. We did not find any results describing how difficult it is to recognise boundaries between particular types of phonemes in literature. Definitely, there is no such information available for Polish as we presented here.

II. Phone Segmentation Method

Phones are characterised by frequency content, so we would expect changes in the power of wavelet resolution levels between phones. Clearly, it would be easier to analyse the absolute value of the rate-of-change of power and expect it to be large at the beginning and at the end of phones. However, this does not uniquely define start and end points for two reasons. First, the power can rise over a considerable length of time at the start of a phone, leading to an ambiguous start time. Second, there may also be rapid changes in power in the middle of a segment. A better method of detecting the boundary of phones relies on power transitions between the discrete wavelet transform (DWT) subbands [3].

In order to obtain DWT, the coefficients of series

$$s(t) = \sum_i c_{m+1,i}\, \phi_{m+1,i}(t) \tag{1}$$

need to be computed, where $\phi_{m+1,i}$ is the ith wavelet function at the $(m+1)$th resolution level. The coefficients of the lower level are calculated by applying the well-known formulae [12]

$$c_{m,n} = \sum_i h_{i-2n}\, c_{m+1,i} \tag{2}$$

$$d_{m,n} = \sum_i g_{i-2n}\, c_{m+1,i} \tag{3}$$

where h and g are the constant coefficients which depend on the assumed pair: scale function ϕ and wavelet ψ. The formulae (2) and (3) are used for the signal decomposition

978-1-4244-5794-6/10/$26.00 ©2010 IEEE

by digital filtering of wavelet coefficients. If the wavelet coefficients $c_{m+1,i}$ of the $(m+1)$th resolution level are given, we can apply (2) and (3) to compute the coefficients of the mth resolution level. The elements of the DWT for a particular level may be collected into a vector, for example $\mathbf{d}_m = (d_{m,1}, d_{m,2}, ...)^T$. The coefficients of other resolution levels are calculated recursively by applying formulae (2) and (3). The multiresolution analysis leads naturally to a hierarchical and fast scheme for the computation of the wavelet coefficients for a given speech signal s. In this way the values

$$\mathrm{DWT}(s) = \{\mathbf{d}_M, \mathbf{d}_{M-1}, ..., \mathbf{d}_1, \mathbf{c}_1\} \quad (4)$$

of the DWT for $M+1$ levels are obtained. The wavelet spectra are produced by using a filter bank (cascading the filtering and downsampling operations). The wavelet transformation can be viewed as a tree. The root of the tree consists of the coefficients of wavelet series (1) of the original speech signal. The next level of the tree is the result of one step of the DWT. Subsequent levels in the tree are constructed by recursively applying the wavelet transform step to split the signal into the low (approximation) and high (detail) parts. The undertaken experiments showed that the speech signal should be decomposed into six levels, which cover the frequency band of a human voice. The energy of the speech signal above 5.5 kHz and below 86 Hz is very low.

The amount $2^{-M+n-1}N$ of wavelet spectrum samples in n-level (where $n = 1, \ldots, M$) depends on the length N of speech signal in time domain, assuming N is a power of 2. For each n-level decomposition the power waveform

$$p_n(i) = \sum_{j=1}^{2^{n-1}} d^2_{n,j+2^{n-1}i} \quad \text{where } i = 0, \ldots, 2^{-M}N - 1, \quad (5)$$

is computed in a different way to obtain the equal number of power samples.

The basic algorithm [3] consists of following steps:

1) Normalise a speech signal by dividing by its maximum value.
2) Decompose a signal into six levels of the DWT.
3) Calculate the sum of power samples in all frequency sub-bands to obtain (5), the power representations $p_n(i)$ of the nth subband.
4) Calculate the envelopes p'_n for power fluctuations in each subband by choosing the highest values of p_n in a window of a given size ω.
5) Calculate the rate-of-change function $r_n(i)$ by filtering $p_n(i)$ with the [1, 2, -2, -1] mask.
6) Given a threshold p of the distance between $r_n(i)$ and p'_n and a threshold $p_{\min}$ of minimal p'_n, find indexes for which $|\beta|r_n(i)| - p'_n(i)| < p$ *AND* $(|\beta|r_n(i+1)| - p'_n(i+1)| > p$ *OR* $|\beta|r_n(i-1)| - p'_n(i-1)| > p)$ *AND* $p'_n(i) > p_{\min}$, where $\beta = 1$. Write such indexes in one vector.
7) Find and group indexes where there is no space between neighbouring ones longer than attribute α.
8) Calculate an average index value (rounded to the nearest integer) for each group found in the previous step as the representative of a group. They are indexes of phones' boundaries in indexing of DWT level 1.

III. Phone Types of Transitions

Detected boundaries may have various degrees of accuracy with respect to hand-segmentation of speech. There are a number of factors that must be considered, including the accuracy of hand-segmented boundaries, since hand-segmentation is not in itself an entirely accurate process because of uncertainties in human perception of the phone boundaries. Additionally, overlapping phones or partially merged phones are a natural phenomena. There is, therefore, a degree of uncertainty in the precision of the boundaries of the phones.

Simply assigning a Boolean value (correct or incorrect) is not really a sensitive measure of segmentation quality. For this reason, fuzzy logic was used, which produces a graded rating of boundary locations in a more sensitive and human-like way. The concept of fuzzy sets and logic was used to derive recall and precision scores. The reasons for why we believe this evaluation is better than the typical ones are presented in [11].

The evaluation was conducted on CORPORA, created under the supervision of Stefan Grocholewski from the Institute of Computer Science, Poznań University of Technology in 1997 [13]. Speech files in CORPORA were recorded with the sampling frequency $f_0 = 16$ kHz, equivalent to sampling period $t_0 = 62.5\mu$s. Speech was recorded in an office with a working computer in the background, which makes the corpus not perfectly clean. Signal to Noise Ratio (SNR) is not stated in the description of the corpus. It can be assumed that SNR is very high for actual speech, but minor noise is detectable for periods of silence.

The database contains 365 utterances (33 single letters, 10 digits, 200 names, 8 simple computer commands, and 114 short sentences), each spoken by 11 females, 28 males, and 6 children (45 people), giving 16 425 utterances in total. One set spoken by a male and one by a female were hand-segmented. The rest were segmented by the dynamic programming algorithm, which was trained on hand-segmented ones.

There are following types of phones in Polish [2]:

1) Stops (/p/, /b/, /t/, /d/, /k/, /g/)
2) Nasal consonants(/m/, /n/, /ni/, /N/)
3) Mouth vowels (/i/, /y/, /e/, /a/, /o/, /u/)
4) Nasal vowels (/e_/, /a_/)
5) Palatal consonants (Pol. Glajdy)(/j/, /l_/)
6) Unstables (Pol. Płynne)(/l/, /r/)
7) Fricatives (/w/, /f/, /h/, /z/, /s/, /zi/, /si/, /rz/, /sz/)
8) Closed fricatives (/dz/, /c/, /dzi/, /ci/, /drz/, /cz/)
9) Silence in the beginnings and ends of recordings
10) Silence inside words

Tables I, II, III and Fig. 1 present evaluation of phone segmentation regarding to transitions of types listed above. Value 0 means that there was no transition of this type.

IV. Discussion

All periods of silence before speech are marked as perfectly detected due to an evaluation algorithm. Apart from that the

TABLE I

RECALL FOR DIFFERENT TYPES OF PHONE TRANSITIONS.

Type	1	2	3	4	5	6	7	8	9	10
1	0.7204	0.6101	0.5114	0.5776	0.5818	0.5007	0.5877	0.6456	0.5210	0.4194
2	0.6015	0.5555	0.4686	0.5812	0.5474	0.5087	0.5817	0.6062	0.5658	0.2129
3	0.4886	0.4493	0.5069	0.0821	0.4605	0.3776	0.4218	0.5872	0.4741	0.3712
4	0.5089	0.4816	0.5384	0	0.4215	0.4388	0.4380	0.5015	0.3692	0.2155
5	0.6403	0.5790	0.4534	0.5362	0.5942	0.5520	0.5829	0.6072	0.5563	0.0702
6	0.5624	0.5445	0.4690	0.5553	0.5428	0.4768	0.5781	0.5558	0.5885	0.2630
7	0.6148	0.5320	0.4389	0.5299	0.4641	0.4708	0.5203	0.5911	0.5784	0.4661
8	0.6216	0.5593	0.4771	0.5424	0.4281	0.5288	0.5372	0.6387	0.5169	0.1388
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0	0.0227
10	0.0399	0.1399	0.4180	0	0.0335	0.0643	0.0835	0.0289	0	0

TABLE II

PRECISION FOR DIFFERENT TYPES OF PHONE TRANSITIONS.

Type	1	2	3	4	5	6	7	8	9	10
1	0.6927	0.5788	0.4783	0.5299	0.5465	0.4741	0.5599	0.6094	0.3115	0.4108
2	0.5523	0.4858	0.4021	0.4996	0.4952	0.4783	0.5375	0.5569	0.3928	0.2129
3	0.4171	0.3692	0.4433	0.0771	0.3963	0.3033	0.3470	0.5207	0.2899	0.3423
4	0.4199	0.4124	0.4735	0	0.3789	0.4073	0.3405	0.4222	0.1987	0.1826
5	0.5943	0.5465	0.3731	0.4688	0.5554	0.5252	0.5488	0.5443	0.3811	0.0645
6	0.4838	0.4976	0.3987	0.4811	0.4811	0.4271	0.5303	0.5174	0.4203	0.2630
7	0.5762	0.4875	0.3732	0.4835	0.4150	0.4208	0.4798	0.5324	0.4158	0.4452
8	0.5573	0.4938	0.4154	0.4926	0.3511	0.4869	0.4809	0.5692	0.3209	0.1333
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0	0.0227
10	0.0365	0.1399	0.4035	0	0.0310	0.0620	0.0835	0.0289	0	0

TABLE III

F-SCORE FOR DIFFERENT TYPES OF PHONE TRANSITIONS. THE SCORES ABOVE 0.5 WERE BOLDED.

Type	1	2	3	4	5	6	7	8	9	10
1	**0.7063**	**0.5940**	0.4943	**0.5528**	**0.5636**	0.4870	**0.5734**	**0.6270**	0.3899	0.4150
2	**0.5759**	**0.5183**	0.4328	**0.5373**	**0.5200**	0.4931	**0.5587**	**0.5805**	0.4637	0.2129
3	0.4500	0.4053	0.4730	0.0795	0.4260	0.3364	0.3807	**0.5519**	0.3598	0.3562
4	0.4601	0.4443	**0.5038**	0	0.3991	0.4225	0.3831	0.4584	0.2583	0.1977
5	**0.6164**	**0.5623**	0.4093	**0.5002**	**0.5742**	**0.5383**	**0.5654**	**0.5740**	0.4523	0.0672
6	**0.5202**	**0.5200**	0.4310	**0.5155**	**0.5101**	0.4506	**0.5532**	**0.5359**	0.4904	0.2630
7	**0.5949**	**0.5088**	0.4034	**0.5056**	0.4382	0.4444	0.4992	**0.5602**	0.4838	0.4555
8	**0.5877**	**0.5245**	0.4441	**0.5163**	0.3858	**0.5070**	**0.5075**	**0.6019**	0.3960	0.1360
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0	0.0227
10	0.0382	0.1399	0.4106	0	0.0322	0.0632	0.0835	0.0289	0	0

periods of silence were not detected very well. The reason for that is that the segmentation method is tuned to phone boundaries and not speech-silence transitions. There are other very efficient methods for this task already established [14].

DWT was also tested for parametrisation of speech [15] and the unvoiced stops (/p/, /t/, /k/) were found more difficult to be recognised than vowels (/aa/, /ax/, /iy/) and unvoiced fricatives (/p/, /t/, /k/). In our case stops were not a big problem for locating them correctly. Actually the highest F-score (0.7063) was for boundaries between two stops and the second grade (0.6270) between stops and closed fricatives. Also transitions from palatal consonants to stops were evaluated highly (0.6164). Transitions between two closed fricatives were another group of easy ones to be detected.

The most difficult for detection were transitions from mouth vowels to any type apart from closed fricatives, especially to nasal vowels (0.0795), unstables (0.3364) and fricatives (0.3807). Also transitions to mouth vowels were difficult to locate correctly. The only exception was from nasal vowels to mouth vowels (0.5038) which is surprisingly much comparing to 0.0795 for a transition in the other way. Another group of boundaries with low F-scores were transitions from nasal vowels apart from the mentioned transition to mouth vowels. Especially difficult were transitions to fricatives (0.3831) and palatal consonants (0.3991). There are no transitions from one nasal vowel into another one. The transitions from closed fricatives to palatal consonants, from unstables to unstables and fricatives to palatal consonants, unstables and another fricatives were also difficult to be detected properly.

According to our results it is relatively easy to find a boundary between phones of the same group if such transition is possible. F-score for such boundaries is usually above 0.5. This is slightly surprising because phones of the same group have typically similar spectra and it could be expected to be difficult to differentiate them.

Tables I, II and III are not symmetric. It is not very

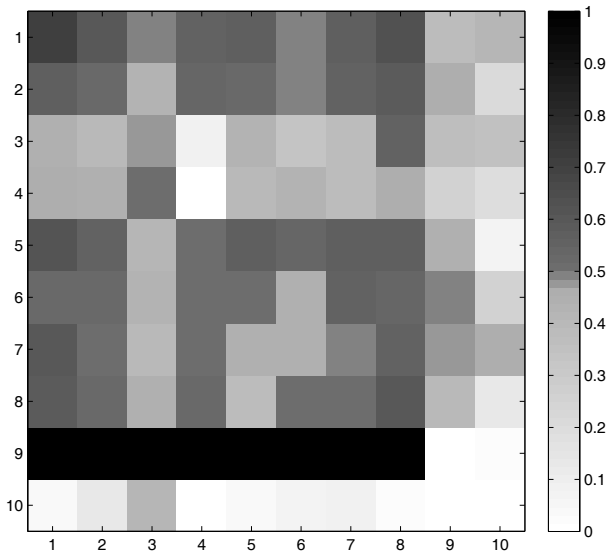

Fig. 1. *F-score of phone boundaries detection for transitions between several types of phones.*

surprising because phone spectra are not symmetric. Their ends and starts can vary significantly. This is why, it might be easier to locate a beginning of a particular phone than its end.

The gained statistical knowledge can be applied. In case of LVCSR, the recognition follows the segmentation. If a phone which is known to cause errors for segmentation is detected, its boundaries can be re-evaluated by another more sophisticated or simply other method. Then another segmentation decision can be taken, leading to a better final recognition.

V. Conclusions

Accurate segmentation is an important task in speech recognition. There are types of phone transitions which are more difficult to detect than others. The average F-score for our segmentation method based on DWT vary from 0.0795 to 0.7063 for transitions between different acoustic types of phones. The experiments support a hypothesis that, in general, it is more difficult to locate boundaries of all vowels than other phones.

VI. Acknowledgements

This work was supported by MNISW grant OR00001905.

References

[1] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, vol. 17, pp. 137–152, 2003.

[2] M. Kępiński, *Kontekstowe związki cech w sygnale mowy polskiej (Eng. Contextual feature relations in Polish speech signal), PhD Thesis*. AGH University of Science and Technology, Kraków, 2005.

[3] B. Ziółko, S. Manandhar, R. C. Wilson, and M. Ziółko, "Wavelet method of speech segmentation," *Proceedings of 14th European Signal Processing Conference EUSIPCO, Florence*, 2006.

[4] K. Stöber and W. Hess, "Additional use of phoneme duration hypotheses in automatic speech segmentation," *Proceedings of ICSLP, Sydney*, pp. 1595–1598, 1998.

[5] D. B. Grayden and M. S. Scordilis, "Phonemic segmentation of fluent speech," *Proceedings of ICASSP, Adelaide*, pp. 73–76, 1994.

[6] C. J. Weinstein, S. S. McCandless, L. F. Mondshein, and V. W. Zue, "A system for acoustic-phonetic analysis of continuous speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 54–67, 1975.

[7] V. W. Zue, "The use of speech knowledge in automatic speech recognition," *Proceedings of the IEEE*, vol. 73, pp. 1602–1615, 1985.

[8] D. Toledano, L. Gómez, and L. Grande, "Automatic phonetic segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.

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

[10] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[11] B. Ziółko, S. Manandhar, and R. Wilson, "Fuzzy recall and precision for speech segmentation evaluation," *Proceedings of 3rd Language and Technology Conference, Poznań*, 2007.

[12] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, pp. 11–38, 1991.

[13] S. Grocholewski, "Założenia akustycznej bazy danych dla języka polskiego na nośniku cd rom (Eng. Assumptions of acoustic database for Polish language)," *Mat. I KK: Głosowa komunikacja człowiek-komputer, Wrocław*, pp. 177–180, 1995.

[14] C. Zheng and Y. Yan, "Fusion based speech segmentation in DARPA SPINE2 task," *Proceedings of ICASSP, Montreal*, pp. I–885–888, 2004.

[15] 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.