Edit Distance Comparison Confidence Measure for Speech Recognition

Dawid Skurzok and Bartosz Ziółko

Abstract A new possible confidence measure for automatic speech recognition is presented along with results of tests where they were applied. A classical method based on comparing the strongest hypotheses with an average of a few next hypotheses was used as a ground truth. Details of our own method based on comparison of edit distances are depicted with results of tests. It was found useful for spoken dialogue system as a module asking to repeat a phrase or declaring that it was not recognised. The method was designed for Polish language, which is morphologically rich.

Keywords Speech recognition decisions · Polish

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) commercial software for continuous Polish. Polish speech contains high frequency phones (fricatives and plosives) and the language is highly inflected and non-positional.

D. Skurzok (✉) · B. Ziółko
Department of Electronics, AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: skurzok@agh.edu.pl.
URL: www.dsp.agh.edu.pl

B. Ziółko
e-mail: bziolko@agh.edu.pl.
URL: www.dsp.agh.edu.pl

J. J. (Jong Hyuk) Park et al. (eds.), *Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 240, DOI: 10.1007/978-94-007-6738-6_19,

It is crucial in a spoken dialogue system to not only provide a hypothesis of what was spoken but also to evaluate how likely it is. A simple probability is not always a good measure because its value depends on too many conditions. In case of dialogue systems, additional measure evaluating if the recognition is creditable or not is very useful. A relation to other, non-first hypothesis can provide it. It allows to repeat a question by a spoken dialogue system or choose a default answer for an unknown utterance. The purpose of Confidence Measures (CMs) is to estimate the quality of a result. In speech recognition, confidence measures are applied in various manners.

Existing types and applications of CMs were well summarised [1–3]. CMs can help to decide to keep or reject a hypothesis in keyword spotting applications. They can be also useful in detecting out-of-vocabulary words to not confuse them with some similar vocabulary words. Moreover, for acoustic adaptation, CM can help to select the reliable phonemes, words or even sentences, namely those with a high confidence score. They can be also used for the unsupervised training of acoustic models or to lead a dialogue in an automatic call-centre or information point in order to require a confirmation only for words with a low confidence score. Recently, applying Bayes based CM for reinforced learning was also tested [4]. A CM based on comparison of phonetic substrings was also described [5]. CMs were also applied in a new third-party error detection system [6]. CMs are even more important in speaker recognition. A method based on expected log-likelihood ratio was recently tested in speaker verification [7]. CMs can be classified [2] according to the criteria which they are based on: hypothesis density, likelihood ratio, semantic, language syntax analysis, acoustic stability, duration, lattice-based posterior probability.

2 Literature Review

Results and views on CMs for speech recognition found in the latest papers were analysed while we worked on our method. In some scenarios it is very important to compute CMs without waiting for the end of the audio stream [2]. The frame-synchronous ones can be computed as soon as a frame is processed by the recognition system and are based on a likelihood ratio. They are based on the same computation pattern: a likelihood ratio between the word for which we want to evaluate the confidence and the competing words found within the word graph. A relaxation rate to have a more flexible selection of competing words was introduced.

Introducing a relaxation rate to select competing words implies managing multiple occurrences of the same word with close beginning and ending times. The situation can be solved in two ways. A summation method adds up the likelihood of every occurence of the current word and adds up the likelihood of every occurence of the competing words. A maximisation method keeps only the occurence with the maximal acoustic score.

The frame-synchronous measures were implemented in three ways regarding a context: unigram, bigram and trigram. The trigram one gave the best results on a test corpus.

The local measures estimate a local posterior probability in the vicinity of the word to analyse. They can use data slightly posterior to the current word. However, this data is limited to the local neighbourhood of this word and the confidence estimation does not need the recognition of the whole sentence. Local measures gave better results on a test set.

Two *n*-gram CMs based evaluations were also recently tested [8] 7-gram based on part-of-speech (POS) tags and 4-gram based on words. The latter was not succesful in detecting wrong recognitions. Applying POS tags in a CM was efficient, probably because it enables analysis on a larger time scale (7-gram instead of 4-gram).

A new CM based on phonetic distance was described [9]. It uses distances between subword units and density comparison (called anti-model by authors). The method employs separate phonetic similarity knowledge for vowels and consonants, resulting in more reliable performance. Phonetic similarities between a particular subword model and the remaining models are identified using training data

$$P\left(X^{\{i\}}|\lambda_{i,1}\right) \geq P\left(X^{\{i\}}|\lambda_{i,2}\right) \geq \cdots \geq P\left(X^{\{i\}}|\lambda_{i,M}\right) \tag{1}$$

where $X^{\{i\}}$ is a collection of training data labeled as model λ_i and $\lambda_{i,m}$ indicates the mth similar model among M subword models compared to the pivotal model λ_i.

Applying of conditional random fields was recently tested [10] for confidence estimation. They allow comparison of features from several sources, namely lattice arc posterior ratio, lattice-based acoustic stability and Levenshtein alignment feature.

3 1-to-3 Comparison

The most widely known CM is based on hypothesis density. It compares the strongest hypothesis with an average of the following n weaker ones (Fig. 1). In our experiments $n = 3$ was empirically found useful and it is a common value for this parameter in other systems as well. Our evaluations were done for sentence error rate. In the first evaluations it worked very well but later on, we found out, that its usefulness is limited in real dialogue applications because it had similar ratio for sentences allowed by a dictionary as for the ones which were not allowed. It was confirmed in later statistical tests with larger dictionaries. This is why we searched for a method based on edit distance comparison and earlier on phonetic substrings [5].

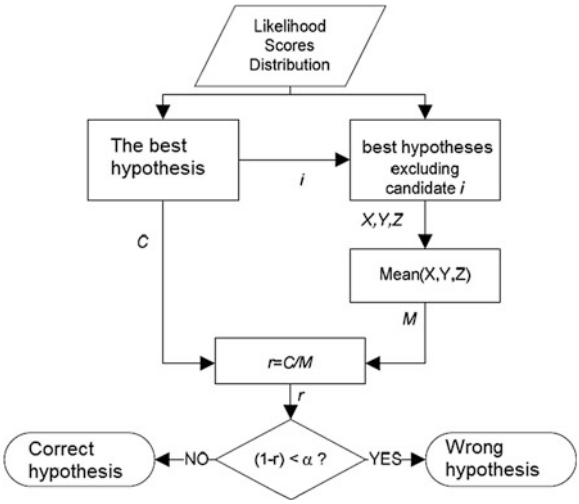

Fig. 1 Algorithm of a standard method of CM by analysis of hypotheses density

4 Edit Distance Comparison

Edit distance comparison CM was designed and implemented for scenarios where there are several utterances very similar to each other. Such situation is especially common in morphologically rich languages like Polish [11], Czech [12] or Finnish [13]. In this type of scenarios classical CMs frequently fail to help detect wrong recognitions. Our new approach operates by measuring Levenshtein distance [14] in phonetic domain between the strongest hypothesis and the following ones. In this method, the mean of edit distances between the first hypothesis and m following ones is taken as the confidence value. We found that $m = 6$ gives the best results (Fig. 2).

Considering only the mean of distances as the confidence indicator, gives worse results then simple 1-to-3 probability comparison, although both methods can be connected to improve final results. Both methods returns numbers from different range and with different variance. We suggest a following formula as a hybrid approach

$$C = C_{1\,to\,3} + \alpha \bar{D}^{\beta}, \tag{2}$$

where C is a final confidence, $C_{1\,to\,3}$ is a confidence calculated using previous method and $\bar{D}$ is a mean of edit distances between the strongest and m following hypothesis. Coefficients α and β are used to scale distance confident and were chosen through optimization. We found that $\alpha = 0.8$ and $\beta = -2$ give the best results.

As it can be concluded, the suggested edit distance comparison method is quite a new approach, which does not fall directly into any of the CM types presented above and listed in literature [2] (Table 1).

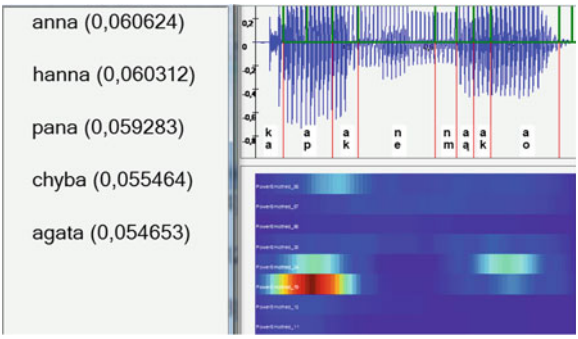

Fig. 2 A screenshot of the developer version of our ASR SARMATA system presents an example of how the described edit distance CM can be applied. The *left* part shows the ranking of top 5 hypotheses and the right one, the time and frequency representation of the analysed audio file. The first three hypothesis have small edit distance between them and the recognition is actually correct

Table 1 Example of calculation of edit distance CM

	Hypothesis	Likelyhood	Distance
1	*/anna/*	0.120	0
2	*/xanna/*	0.095	1
3	*/panna/*	0.080	1
4	*/pana/*	0.065	2

For this case, let us assume that $m = 3$. The 1-to-3 confidence is C_1 to $_3 = 0.12/((0.095 + 0.08 + 0.065)/3) = 0.12/0.08 = 1.5$. The hybrid confidence (2) is $C = 1.5 + 0.8((1 + 1+2)/3)^{-2} = 1.5 + 0.8 \cdot 1.33^{-2} = 1.5 + 0.45 = 1.95$

5 Tests and Results

The standard 1-to-3 method was compared with the edit distance method in a sequence of experiments on as test corpus based on CORPORA [15]. The recordings consists of 4435 audio files, each with one word spoken by various male speakers. The audio files have sampling rate 16 kHz and 16-bit rate. No language model was used in the tests. Some of the words in test corpora were recorded as isolated word, while others were extracted from longer sentences. All tests were made using SARMATA ASR system [11]. The dictionary has 9177 words. 1492 of total 4435 words were recognised correctly (Table 2).

Table 2 Result for different methods ED is an abbreviation of edit distance confidence

	1-to-3	ED	1-to-3 + ED
Precision	0.71	0.38	0.72
Recall	0.65	0.77	0.65
Accuracy	0.79	0.50	0.80
F-measure	0.68	0.51	0.70

6 Conclusions

The suggested CM method based on edit distance enhanced the classical 1-to-3 method in an experiment motivated by real applications and end-user tests. The method was designed for morphologically rich languages, like Polish, as it gives better scores if the strongest hypotheses are phonetically similar. The presented method gives 2 % improvement in F-measure and 1 % improvement in accuracy.

Acknowledgments The project was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/D/ST6/00914.

References

1. Guo G, Huang C, Jiang H, Wang RH (2004) A comparative study on various confidence measures in large vocabulary speech recognition. Proceedings of international symposium on Chinese spoken language, pp 9–12
2. Razik J, Mella O, Fohr D, Haton J (2011) Frame-synchronous and local confidence measures for automatic speech recognition. Int J Pattern Recognit Artif Intell 25:157–182
3. Wessel F, Schluter R, Macherey K, Ney H (2001) Confidence measures for large vocabulary continuous speech recognition. IEEE Trans Speech Audio Proc 9(3):288–298
4. Molina C, Yoma N, Huenupan F, Garreton C, Wuth J (2010) Maximum entropy-based reinforcement learning using a condense measure in speech recognition for telephone speech. IEEE Trans Audio, Speech Lang Proc 18(5):1041–1052
5. Ziółko B, Jadczyk T, Skurzok D, Ziółko M (2012) Confidence measure by substring comparison for automatic speech recognition. ICALIP, Shanghai
6. Zhou L, Shi Y, Sears A (2010) Third-party error detection support mechanisms for dictation speech recognition. Interact Comput 22:375–388
7. Vogt R, Sridharan S, Mason M (2010) Making confident speaker verification decisions with minimal speech. IEEE Trans Audio Speech Lang Process 18(6):1182–1192
8. Huet S, Gravier G, Sebillot P (2010) Morpho-syntactic post-processing of n-best lists for improved French automatic speech recognition. Comput Speech Lang 24:663–684
9. Kim W, Hansen J (2010) Phonetic distance based condense measure. IEEE Signal Process Lett 17(2):121–124
10. Seigel M, Woodland P (2011) Combining information sources for confidence estimation with crf models. Proceedings of InterSpeech
11. Ziółko M, Gałka J, Ziółko B, Jadczyk T, Skurzok D, Mąsior M (2011) Automatic speech recognition system dedicated for Polish. Proceedings of Interspeech, Florence
12. Nouza J, Zdansky J, David P, Cerva P, Kolorenc J, Nejedlova D (2005) Fully automated system for Czech spoken broadcast transcription with very large (300 k+) lexicon. Proceedings of InterSpeech, pp 1681–1684
13. Hirsimaki T, Pylkkonen J, Kurimo M (2009) Importance of high-order n-gram models in morph-based speech recognition. IEEE Trans Audio Speech Lang Process 17(4):724–732
14. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Phys Doklady 10:707–710
15. Grocholewski S (1998) First database for spoken Polish. Proceedings of international conference on language resources and evaluation, Grenada, pp 1059–1062