Inter-Annotator Agreement on Spontaneous Czech Language

Limits of Automatic Speech Recognition Accuracy

Tomáš Valenta, Luboš Šmídl, Jan Švec, and Daniel Soutner

University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 22, 306 14 Plzeň, Czech Republic
{valentat, smidl, honzas, dsoutner}@kky.zcu.cz

Abstract. The goal of this article is to show that for some tasks in automatic speech recognition (ASR), especially for recognition of spontaneous telephony speech, the reference annotation differs substantially among human annotators and thus sets the upper bound of the ASR accuracy. In this paper, we focus on the evaluation of the inter-annotator agreement (IAA) and ASR accuracy in the context of imperfect IAA. We evaluated it using a part of our Czech Switchboard-like spontaneous speech corpus called Toll-free calls. This data set was annotated by three different annotators rendering three parallel transcriptions. The results give us additional insights for understanding the ASR accuracy.

Keywords: automatic speech recognition, inter-annotator agreement, accuracy.

1 Introduction

Automatic speech recognition accuracy differs significantly among various domains and tasks. On particular tasks in certain domains, the recognition accuracy almost achieves 100 %, whereas in others, it is about 60 % or less, see Fig. 1 [1].

A similar trend can be observed for human transcriptions when applying the inter-annotator agreement (IAA). It can be almost perfect in domains like dictation of written text, whereas human transcriptions of spontaneous spoken language may differ substantially. The estimate of human transcription error in Fig. 1 is quite optimistic and does not take into account the domain.

The inter-annotator agreement of classification tasks such as assigning labels from few classes (e.g. parts-of-speech tagging, named entity recognition etc.) is commonly evaluated. On text, the IAA is evaluated in transcription and translation of historical texts [10] or psychoanalysis, and sometimes in speech synthesis [12]. Surprisingly, the IAA is not commonly evaluated in the area of speech recognition. The transcription(s) available are viewed as a gold-standard no matter how accurate they are.

For the evaluation of the IAA in classification tasks assigning few labels only, statistics compensating the chance agreement are often used, e.g. Cohen's κ [2]. The chance agreement in ASR tasks is, however, inversely proportional to the vocabulary size, hence it has virtually no effect on the value of the IAA.

The goal of this article is to show that on some tasks, human transcription accuracy (i.e. IAA) above 90 % is almost unachievable. By implication, this sets the upper bound for the automatic speech recognition accuracy far below 100 %.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 390–397, 2014.

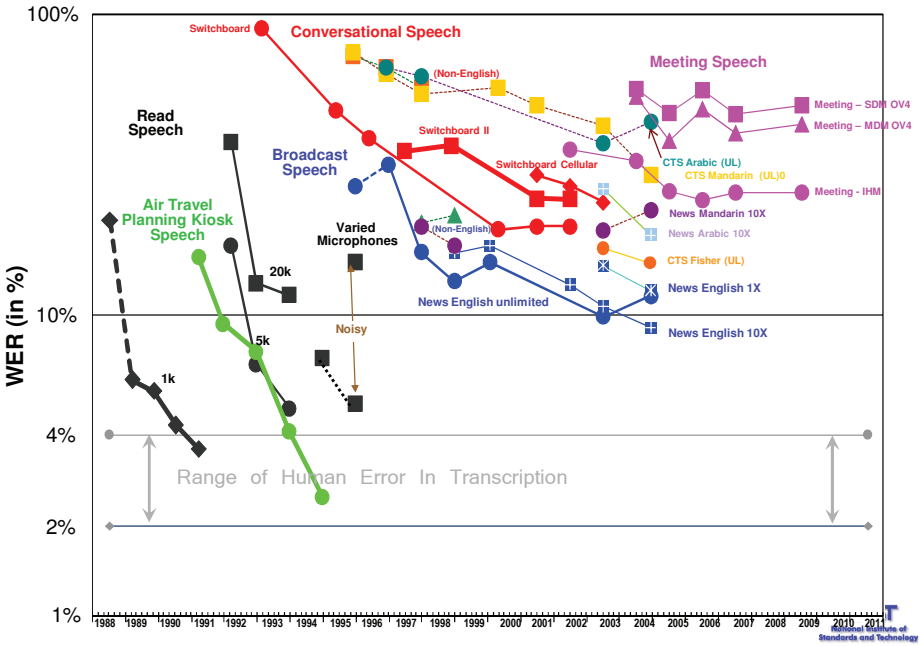

Fig. 1. National Institute of Standards and Technology speech-to-text benchmark history — May 2009 [1]. Word error rate used as opposed to the accuracy used in the rest of the article.

To demonstrate a lower IAA and its impact on the recognition accuracy, a Czech Switchboard-like corpus was chosen. It contains recordings of telephone communication of two people. The callers usually know each other very well, they use lots of non-standard or local words and they speak colloquially. This reduces the recognition performance significantly as well as the ability to recognize and understand the conversation by other people.

In Section 2, the evaluation corpus is described. All results, including the human and automatic speech recognizer, were computed on evaluation sets of the corpus. In Section 3, the inter-annotator agreement methods are described and the results obtained are presented. In Section 4, the IAA results are compared to an automatic speech recognition system. Finally, Section 5 summarizes the results and draws some conclusions.

2 The Corpus

The evaluation corpus is called *Toll-free Calls*. It is similar to the standard Switchboard corpus [3]. It consists of many hours of spontaneous spoken Czech language used in phone calls.

To record the corpus, a simple dialogue system was developed. When the call was received, it was rejected and a few seconds later the system called back the caller. As nothing was charged to the caller, hence *Toll-free Calls*. Then the system asked the caller

```
Czech: <ehm_ANO> <unintelligible> (nějak(ňák)) zvláštně
English: <erm_YES> <unintelligible> (some(sum)) weirdly
```

Fig. 2. Annotation example. Non-speech events are in angle brackets, round brackets mark orthographic transcription and what was actually pronounced.

for a phone number he/she wanted to dial and as soon as the call was connected, both call participants were notified that their conversation would be recorded and used for research purposes. Both parties could talk for up to 15 minutes.

The audio was recorded in a standard telephone quality, i.e. 16bit PCM, frequency 8 kHz, stereo — one channel for the caller and the other for the person receiving the call. For the annotation, the channels were split and each was annotated separately to protect the call participants' privacy. Each channel was subsequently segmented to utterances which were then transcribed.

A major part of the corpus was transcribed by a single annotator to obtain training data for the ASR system. The balance of the data (evaluation part) was transcribed by several annotators allowing us to calculate the IAA.

The transcriptions were made by well-experienced annotators with focus on acoustic modelling which means that the transcriptions were not orthographic and several non-speech event tags, exact pronunciations etc. were used in the annotation, see Fig. 2 for an example.

2.1 Evaluation Sets

The part of the corpus used for evaluation was annotated by several annotators. Each sentence was annotated three times by different annotators. Two evaluation sets, *Set1* and *Set2*, were selected so that the number of sentences covered by three same annotators was maximal. The sets contain 2,003 and 2,838 sentences respectively. Set1 was annotated by annotators A, B and C, Set2 was annotated by annotators A, B and D.

Prior to the evaluation, the text was pre-processed in the following way:

- non-speech events were removed;
- converted to lower-case;
- punctuation removed (replaced by spaces);
- exact pronunciations removed (orthographic transcription used if available, see Fig. 2);
- multiwords (e.g. good_morning) were split;
- white-space normalized;
- commonly confused words (frequency greater than two) were considered as equivalent.

In spoken Czech, some words pronounced similarly (*sem*, *jsem*; *byli*, *byly*), orthographically vs. colloquially (*mladý*, *mladej*) or just typing errors were fixed this way.

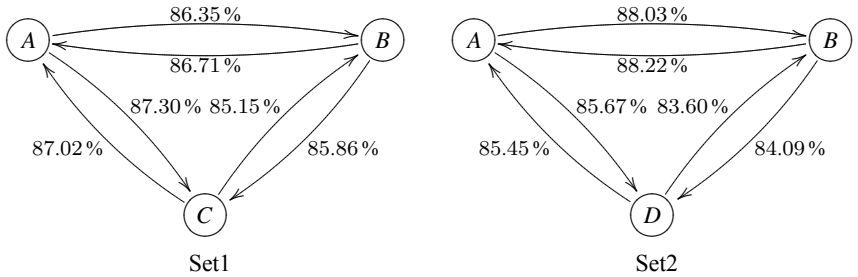

Fig. 3. Inter-annotator agreement on evaluation sets. Annotator at the origin of the arrow was used as a reference and at the tip as a (recognition) hypothesis.

Table 1. Average inter-annotator agreement on evaluation sets.

Set1

	A	B
B	86.53 %	
C	87.16 %	85.51 %

Set2

	A	B
B	88.13 %	
D	85.56 %	83.85 %

3 Inter-annotator Agreement

The inter-annotator agreement was calculated in the same way as recognition accuracy is calculated. First, the annotations were aligned so that the Levenshtein distance [7] was minimal. Penalties for substitution, insertion and deletion were set to be the same as in HTK toolkit, i.e. $p_S = 10$, $p_I = 7$ and $p_D = 7$ respectively [15]. Accuracy is then defined by (1):

$$Acc = \frac{N - S - I - D}{N}, \tag{1}$$

where N is the number of words in the reference, S is the number of substitutions, I is the number of insertions, and D is the number of deletions.

Being penalties $p_S \neq p_I + p_D$ and annotations (one used as a reference and one as a recognition hypothesis) having different length, the accuracy depends on which annotation is used as reference, as shown in Fig. 3. In Table 1, weighted average, with respect to the number of words in the reference, is taken from the numbers to obtain a simple measure. It gives the same result as if the reference consisted of concatenation $A \| B$ and the hypothesis of $B \| A$.

Averaging the numbers in Table 1 we get

$$IAA_1 = 86.40\,\% \qquad \text{and} \qquad IAA_2 = 85.84\,\% \tag{2}$$

for evaluation sets Set1 and Set2 respectively. If we wished to get an overall estimate of inter-annotator agreement on Toll-free Calls corpus, we can take a weighted average, with respect to the number of sentences, of the two numbers and we get

$$IAA = 86.01\,\%. \tag{3}$$

Table 2. ASR performance: plain accuracy, accuracy after rescoring, oracle accuracy and multi-oracle accuracy.

	Set1			Set2		
Annotator	A	B	C	A	B	D
Accuracy	50.35 %	49.43 %	48.41 %	55.70 %	54.17 %	53.46 %
Rescore-Acc	54.50 %	53.98 %	52.69 %	60.45 %	58.89 %	57.98 %
Ora-Acc	67.73 %	66.66 %	65.56 %	72.60 %	70.57 %	69.69 %
Mul-Ora-Acc		52.41 %			58.07 %	

4 Evaluating an ASR System

The recognition engine follows a standard design [9]. The acoustic models in our system are based on the hidden Markov models (HMM). Standard 3-state left-to-right models with a mixture of 16 Gaussians in each state are used (5000 states totally). The speech data were parametrized as 12-dimensional PLP [5] cepstral features including their delta and delta-delta derivatives. Cepstral mean subtraction was applied per speaker. We used a real-time large vocabulary continuous speech recognizer (LVCSR) to achieve a high degree of interactivity. The LVCSR system [11] uses lexical trees and Viterbi search using 3-gram language models with Witten-Bell discounting [14]. The ASR recognition vocabulary contains 123,038 words and the OOV rate on evaluation data is below 1 %.

In Figure 4 and Table 2 recognition accuracy is calculated against each annotation taken as a reference. Average IAA is put on arcs connecting the annotators.

4.1 Rescoring

We took the word lattices from the first recognition pass and extracted n best hypotheses (with $n = 1,000$) for each utterance; this n-best list was the base for our experiments with language models. The second pass of the recognition was provided with a more complex language model to discover if the results could be improved. After several

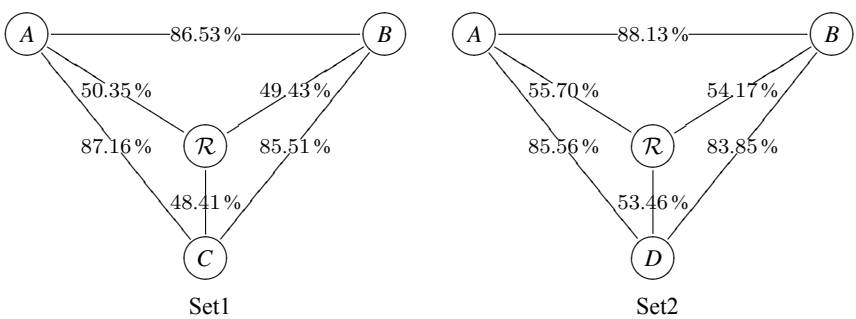

Fig. 4. Recognition accuracy compared to the average inter-annotator agreement, $\mathcal{R}$ stands for the recognizer.

experiments we rescored the n-best list with recurrent neural network language model (RNN LM) [8] in combination with background LM which is a standard 5-gram smoothed with Kneser-Ney discounting [6].

The choice of RNN was motivated by its ability to capture longer context than n-gram models and because of better performance on smaller corpus, which is our case (training data size is 5.4 M words). The vocabulary of RNN LM was shortlisted to 40 k most frequent words, the size of the hidden layer was chosen $h = 400$ and the model was trained together with 8 M maximum entropy features.

Running the second recognition pass described above, the recognition accuracy can be increased slightly, see Table 2.

The upper bound of rescoring performance is given by the oracle accuracy. It is the maximum accuracy of any recognition hypothesis from the n-best list. See Table 2.

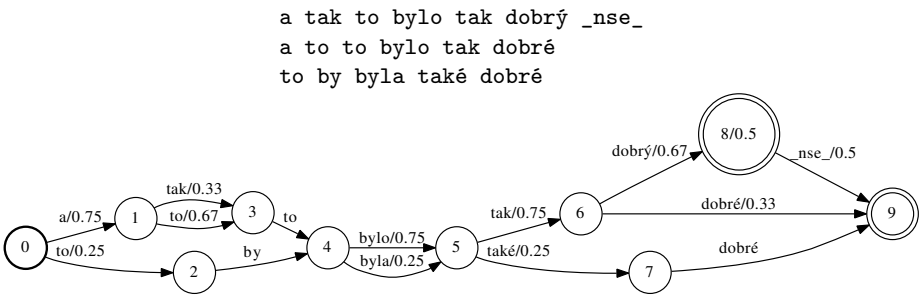

Fig. 5. Multi-oracle accuracy lattice made from three transcriptions. _nse_ means an optional non-speech event.

Having three parallel annotations, we can reverse the oracle accuracy calculation. We take the best hypothesis from the recognizer (not rescored) and try to match it as closely as possible against a lattice created from all annotations for the utterance to get a multi-oracle accuracy. For an example of this type of lattice (generalized confusion network), see Fig. 5. The results are shown in Table 2.

5 Conclusion

In the article we defined the inter-annotator agreement calculation on a speech corpus. We evaluated it on spontaneous speech corpus called Toll-free calls, where it shows to be surprisingly low.

Spoken Czech language differs from its written form significantly, a lot of colloquialisms is used. In spontaneous and expressive [4] communication, lots of non-verbal sounds are also used that are hard to transcribe. Moreover, written form of some words is dependent on the context (gender of the subject) which cannot be told from the speech segments transcribed by the annotators. Most of these phenomenons were, however, mitigated by the text preprocessing and the rest is the very matter of the inter-annotator disagreement, which sets the upper bound to the recognition accuracy.

Unquestionably, the ASR results are worse than human transcription, although it should be noted that the ASR processes the audio in real time in a single pass. In contrast, a human annotator works about 8 times slower than real time, and also has the opportunity to play back the recording repeatedly. The annotator also "recognizes" the spoken data in real time. However transcribing the text with all of the non-speech tags is considerably time-demanding. Undertaking a second recognition pass takes about twice as long as real time and can increase the accuracy. Yet it still fails to meet the IAA.

The oracle accuracy, a measure of how we can increase the accuracy by improving the language model or by rescoring, comes quite close to the IAA. On the other hand, the language modelling of such data is a difficult task, because a lot of non-verbal sounds are used in the communication, the utterances are very short. The number of words used is relatively small, although theoretically unlimited.

The presented results demonstrate that achieving 100% recognition accuracy in spontaneous or expressive speech is impossible and often not necessary. For example, natural language understanding tasks often do not require perfectly recognized utterance (best hypothesis), more efficient techniques (e.g. keyword spotting) can be used [13]. Moreover, most transcription/recognition errors are caused by words with no important sense (hesitation, fillings etc.).

The inter-annotator agreement sets the upper bound that can be achieved in the automatic speech recognition accuracy. In theory, it is possible for the recognition accuracy to go higher, but that would only signal an over-fitting to a particular transcription. For domains with low IAA, it would be suitable to accompany the recognition accuracy reports with the IAA value.

Acknowledgement

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005), is greatly appreciated.

This work was supported by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090. Ministry of Education, Youth and Sports of the Czech Republic project No. LM2010013. The work has been supported by the grant of the University of West Bohemia, project No. SGS-2013-032.

References

1. Ajot, J., Fiscus, J.: Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results. NIST Rich Transcription Evaluation Workshop (2009)
2. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37–46 (Apr 1960)
3. Godfrey, J., Holliman, E., McDaniel, J.: SWITCHBOARD: telephone speech corpus for research and development. In: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 517–520 vol. 1. IEEE (1992)
4. Grůber, M., Tihelka, D.: Expressive speech synthesis for Czech limited domain dialogue system — Basic experiments. In: IEEE 10th International Conference on Signal Processing Proceedings. pp. 561–564. IEEE (Oct 2010)

5. Heřmanský, H.: Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87(4), 1738 (1990)
6. Kneser, R., Ney, H.: Improved backing-off for M-gram language modeling. In: Proceedings of Acoustics, Speech, and Signal Processing, ICASSP-'95, vol. 1, pp. 181–184 (1995)
7. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707 (1966)
8. Mikolov, T., Kombrink, S., Deoras, A., Burget, L., Černocký, J.: RNNLM — Recurrent Neural Network Language Modeling Toolkit. Proceedings of ASRU 2011, pp. 1–4 (2011)
9. Müller, L., Psutka, J., Šmídl, L.: Design of speech recognition engine. In: Sojka, P., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue, Proceedings. Lecture Notes in Artifical Intelligence, vol. 1902, pp. 259–264, Springer-Verlag Berlin, Germany (2000)
10. Munyaradzi, N.: Transcription of the Bleek and Lloyd Collection using the Bossa Volunteer Thinking Framework. Ph.D. thesis, Department of Computer Science, University of Cape Town, Cape Town (2013)
11. Pražák, A., Psutka, J.V., Hoidekr, J., Kanis, J., Müller, L., Psutka, J.: Automatic online subtitling of the Czech parliament meetings. Text, Speech and Dialogue 4188, 501–508 (2006)
12. Tihelka, D., Romportl, J.: Statistical evaluation of reliability of large scale listening tests. In: Proceedings of 9th International Conference on Signal Processing. pp. 631–636. IEEE (Oct 2008)
13. Švec, J., Šmídl, L., Ircing, P.: Hierarchical Discriminative Model for Spoken Language Understanding. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013. pp. 8322–8326. IEEE, Vancouver, Canada (2013)
14. Witten, I.H., Bell, T.C.: The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37(4), 1085–1094 (1991)
15. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK (2006)