End-to-End Recognition of Spontaneous Speech on the Hungarian BEA Database

Tímea Fekete, Péter Mihajlik[1,2]

[1] *Budapest University of Technology and Economics*
[2] *Hungarian Research Centre for Linguistics*

Abstract

The end-to-end deep neural network based speech recognition approach is increasingly popular due to its fully data driven nature - no language-specific knowledge is needed beyond the transcribed speech data. However, most of the end-to-end speech recognition experiments are performed on read speech and no Hungarian language results are available for the Speech Community. In this paper, we make the first attempt to train and evaluate a Hungarian speech recognition system based on the studio-quality Hungarian BEA (Spoken Language Speech Database) in an end-to-end neural manner. We present the challenge of recognising spontaneous speech: even without any significant background noise, the word error rate on spontaneous speech is an order of magnitude higher than in the case of planned speech - both recorded with the same speakers in the same environment. This emphasises the need for more thorough studies of spontaneous speech and possibly for more data.

Keywords: automatic speech recognition, end-to-end neural model, deep learning, Hungarian

1. Introduction

With the spread of artificial neural networks and deep learning, an incredible progress is taking place in numerous fields of technology. Nowadays' best results in automatic speech recognition (ASR) come from end-to-end models, models that are based entirely on deep neural networks. Despite this fact, no previous publication is known to the authors aiming at Hungarian language end-to-end speech recognition on publicly available databases.

As Mihajlik (2020) argues, end-to-end ASR can be considered as a result of natural evolution of ASR systems: statistics and machine learning have been

Email addresses: fe.timea@gmail.com (Tímea Fekete), mihajlik.peter@nytud.hu (Péter Mihajlik)

Beszédtudomány – Speech Science 2021. 261–272. DOI: 10.15775/Besztud.2021.261-272

always essential "ingredients" in an ASR recipe (Bahl et al., 1983). The key step in building the entire ASR on one neural network was the introduction of CTC (Connectionist Temporal Classification) algorithm (Graves et al., 2006), still popular in research and production level ASR systems. This allowed to train acoustic models directly on characters (graphemes), without the need for any pronunciation dictionary or phonological or phonetic knowledge. The approach, however, is mostly applied for isolating languages - like English - where shorter words fit the algorithm more. The effectiveness of the purely CTC-based end-to-end approach is questionable in the case of agglutinating languages where the average word length can be much more. On the other hand, spontaneous speech tends to use shorter words.

Therefore, we decided to train and evaluate an end-to-end deep neural ASR network on the Hungarian language BEA database where we can investigate the challenges of spontaneous speech in contrast to the much more frequent read speech databases.

2. Experiments

One of the crucial elements in the development of automatic speech recognition is data. A publicly available option for English is LibriSpeech, a corpus derived from audiobooks; created specifically for training and evaluating speech recognition systems (Panayotov et al., 2015). It has become one of the most used databases for such purposes, providing a basis of comparison for different models.

BEA is a Hungarian speech corpus developed by the Research Institute for Linguistics in Hungary for various research purposes. This database, unlike LibriSpeech, contains spontaneous speech for the most part, but there are also some other kinds of speech during which the participants repeated sentences or read a text aloud (Gósy et al., 2012) - for simplicity, we will refer to this 'not spontaneous' part of data collectively as 'planned speech'. Most public databases contain read texts exclusively, which makes BEA stand out.

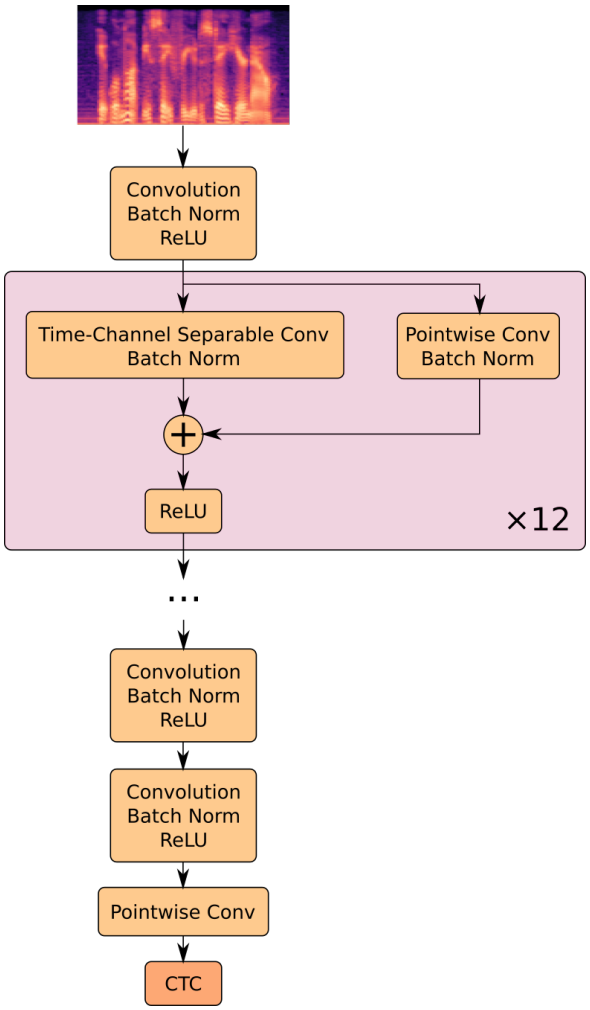

Figure 1: The QuartzNet 12×1 end-to-end deep neural network architecture

We trained an end-to-end ASR model on both LibriSpeech and BEA, enabling a comparison between the two performances. For BEA, the training dataset contained 87.59 hours of speech, the evaluation dataset contained 7.62 hours - distributed similarly to the LibriSpeech dataset in use, which had a 92.34hrs/4.28hrs ratio for the train-clean-100 and test-clean subsets.

We used QuartzNet, a state-of-the-art CTC-based neural acoustic model developed by NVIDIA (Kriman et al., 2019). It is composed of multiple blocks

containing one or more modules with 1D time-separable convolutional layers, batch normalisation and ReLU layers. The blocks connected with residual connections are denoted by B whereas C stands for other (non-residual) blocks. Our model shape is 12×1, which means that there are 12 internal blocks ($B_1 - B_{12}$), each containing one module with residual connection and this internal structure is completed by leading and trailing C-type blocks ($C_1 - C_4$). The total number of parameters are around 5M. For the details, see Figure 1 and Table 1. Although more complex structures are possible with QuartzNet (which also might bring better results), we had to consider resource-efficiency. There are multiple end-to-end architectures using convolutions, such as wav2letter (Pratap et al., 2019) and Jasper (Li et al., 2019). QuartzNet works with fewer parameters - time-channel separable convolutions use less weights (kernel size × input channels + input channels × output channels) than regular 1D channel convolutions (kernel size × input channels × output channels).

Table 1: The main parameters of the consecutive convolutional blocks of the QuartzNet architecture. There are 36 output labels: the 35 Hungarian characters plus the space character

Block	**Kernel size**	**Output channels**
C_1	33	256
$B_1 - B_3$	33	256
$B_4 - B_6$	39	256
$B_7 - B_9$	51	512
$B_{10} - B_{12}$	63	512
C_2	75	512
C_3	1	1024
C_4	1	36 (labels)

To reduce the possibility for overfitting, data augmentation, i.e. adding slightly modified copies of our data was used. In speech recognition, a well-

performing method is SpecAugment, during which three kinds of modifications are applied: time warping (deforming the features through time), frequency channel masking (a block of frequencies removed) and time masking (a block of series removed) (Park et al., 2019).

The minibatch length for training was set to 32 and to 1 for evaluation. A learning rate scheduler was applied with a linear warmup of 5% of training time targeting a maximum value of 0.05, then decreasing dynamically with cosine annealing, keeping a minimum of 0.001. The training lasted 100 epochs. Novograd optimizer (Ginsburg et al., 2020) was used with a weight decay of 10^{-4} and $\beta_1 = 0.95$, $\beta_2 = 0$. During inference, greedy CTC decoding was performed only because we wanted to avoid the application of any external lexicon or language model so that the comparison of planned and spontaneous results are not biased by prior knowledge. All the experiments were performed on an NVIDIA GeForce GTX 1070 GPU.

3. Results

3.1. Quantitative analysis

The most widely used metric for evaluating speech recognition models is WER, standing for Word Error Rate. It shows the rate of errors on a word level using the minimum number of single-word edits as a distance between the reference and hypothesis. The less the value, the better the performance it indicates. Similarly, we can calculate CER (Character Error Rate) on a character level. The later might be even more important for Hungarian, as it's a morphologically rich language.

The training took about three times as many steps for the Hungarian dataset, because it had three times more files than the English one, despite totalling a similar length of time (segments containing only a few words were frequent in the BEA dataset, not so in LibriSpeech). Different results can be expected for the planned and spontaneous parts of the BEA evaluation set, therefore we also evaluated them independently. In Figure 2, we can see how the WER had

changed during the training of the models for both languages (trained with the same parameters).

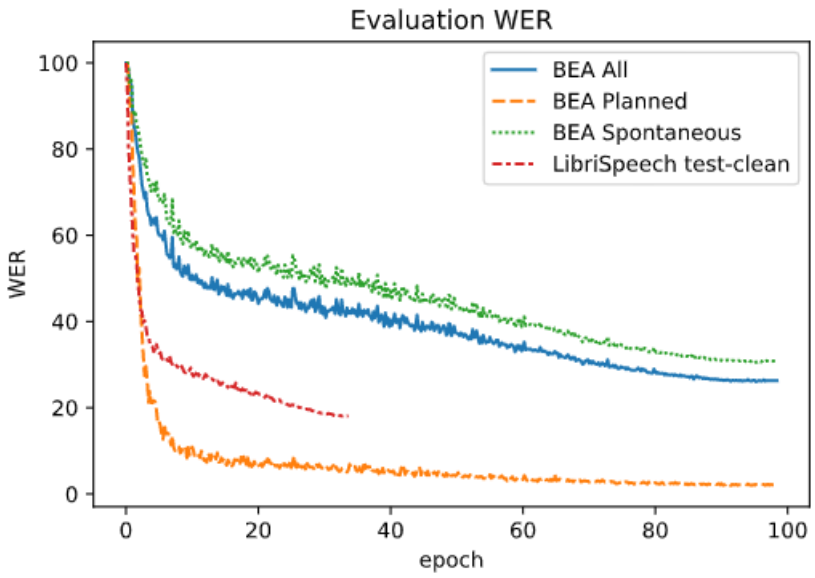

Figure 2: Evaluation WER

Table 2: The model's performance on different evaluation datasets

	Word count	WER	Char count	CER
BEA Planned	7,679	2.09%	63,701	0.46%
BEA Spontaneous	40,561	30.84%	269,384	8.98%
BEA All	48,240	26.28%	333,085	7.35%
LibriSpeech test-clean	49,957	18.16%	281,530	5.81%

The error rates for the planned subset of BEA got quite close to zero (see Table 2 and Figure 3), due to the nature of the data: the same sentences were repeated by different participants, so it can be expected from the model to recognise these particular sentences almost perfectly. On the other hand, it did not perform so well on the spontaneous subset, even compared to the

model trained on LibriSpeech - spontaneity brings obvious difficulty into speech recognition. Our results aren't necessarily on par with the latest state-of-the-art ones, but note that we trained a less complex QuartzNet architecture on a smaller amount of data (only 100 hours instead of the usual 1000 hours) than most reported models.

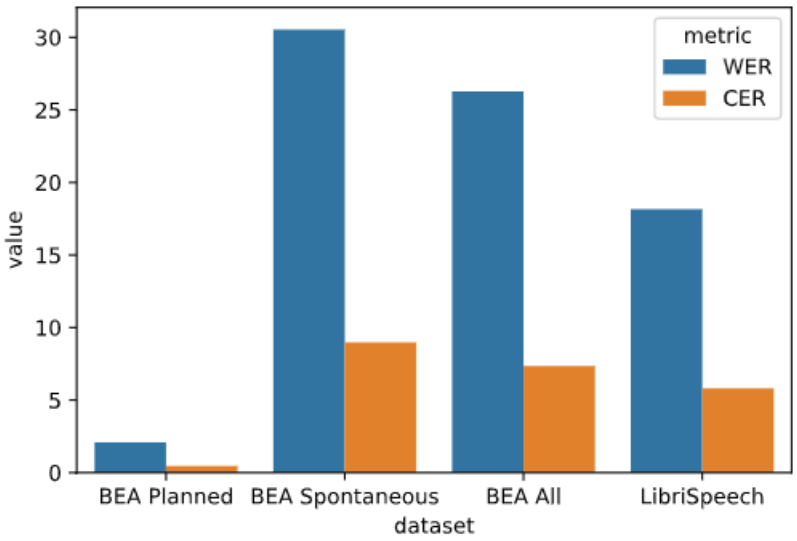

Figure 3: Error rates

3.2. Subjective analysis of spontaneous recognition errors

The magnitude of WER on spontaneous speech looks unacceptably high. However, if we zoom in on some random recognition results on the spontaneous subset, it turns out that many of the errors seem acoustically forgivable - reading aloud some of the hypotheses, one could hardly notice the difference from the reference transcription.

1.	/English/	/but I really like to cook and bake/
	Reference	de nagyon szeretek főzni sütni
	Hypothesis	de nagyon szeretek főzni süt nény

Here, similarly to the English term *chutney*, the pronunciation of *süt nény* can be quite close to *sütni*.

2.	/English/	/and the problem and solution must be sought in the family background as well/
	Reference	és a családi háttérben kell keresni a problémát és a megoldást is
	Hypothesis	és a családi áttérben kell keresni a problémát és a megoldást is

The lack of *h* in this case could go unnoticed in spontaneous situations.

3.	/English/	/roots have been grown where they have been planted with a foreign plant/
	Reference	gyökeret növesztettek ahol idegen növény mellé ültették
	Hypothesis	gyökeret növesztettek ahol idegán növény mellé ültették

In this particular case, the cause of recognition error could be due to an accent typical in the Highland region of former Hungary.

4.	/English/	/it is still easier to accept this way but he has beheaded the citizens/
	Reference	mégiscsak egyszerűbb így elfogadni de a polgárokat lefejeztette
	Hypothesis	mégiscsak egyszerűbb üígy el fog adni de a polgárokat lefejeztette

Now we can observe two types of errors. The first one is explainable with a kind of co-articulation of *egyszerű* and *így*. The second one has a semantically false segmentation but there is not necessarily an acoustical difference between *elfogadni* and *el fog adni*. A possible explanation is that the shorter words might have occurred more frequently in the training set.

5.	/English/	/so his hand was very roughly cut apart because he got it in front of his face/
	Reference	hát nagyon durván szét volt vágva a keze mer az arca elé kapta
	Hypothesis	hát nagyon durván szét volt vágal kezde mer az arc elé kapta

This type of misrecognitions are not easy to explain - we have to admit that the artificial neurons were probably not trained adequately.

6.	/English/	/but if I go down to her let's say she lives very far away just/
	Reference	de hogyha hozzá lemegyek mondjuk ő nagyon messze lakik csak
	Hypothesis	de hogyha hozzále megyek mondjukő nagyon messzel akik csak

An illustrative example where the only source of word errors is false segmentation causing a local WER of 70% , whereas the hypothesis is totally correct on the phonetic level.

7.	/English/	/finish the ice cream and even top with a good deal of chocolate so/
	Reference	befejezni a fagyit és még a tetejére is jó sok csokit tehát így
	Hypothesis	befejezni a pajyit és még a teteére is jó sok csokit tehát így

The *pajyit* hypothesis for *fagyit* is odd but not so impossible - the phonetic representation of both f and p is labial, and the Hungarian speech sound for *gy* is close to the assimilated form of d+j. The dropping of j from *teteére* would be completely tolerable perceptually if this hypothesis was synthesised.

8.	/English/	/and I actually put myself on these pages to practice English/
	Reference	és tulajdonképpen azér tettem föl magam ezekre az oldalakra hogy gyakoroljam az angolt
	Hypothesis	és csak azér tettem föl magam ezekre zudorkho gyakorolan az angolt

It might look bizarre that a longer word (*tulajdonképpen*) was mistaken for a completely different short one (*csak*). In this case, the speaker was speaking

so fast that only trained ears could understand the other highlighted part as well.

Some errors are due to incorrect word segmentation, non-space characters do match in such cases. This shows that the boundaries of words aren't always clear during spontaneous speech. Other times, a few letters were left out or changed, resulting in nonsense words. Such problems could have been avoided using a language model, which was out of scope from this study. Errors on the planned part of test data hardly ever occurred.

Examples from the model trained on LibriSpeech:

1.	Reference	yes all alone by himself asserted jasper vehemently and winking furiously to the others to stop their laughing he did now truly phronsie
	Hypothesis	yes all alone by himself a serted jasper veemently and winking puriously to the others to stop their laughing he did now truly phronsie
2.	Reference	we sat with the officers some little time after dinner and then went ashore
	Hypothesis	we sat wit the officer some little time after dinner and then wen ashore
3.	Reference	from the norwegian graveyard one looks out over a vast checker board marked off in squares of wheat and corn light and dark dark and light
	Hypothesis	from the norregiond grave yard one looks out over a vast checkerbord marked off end squares of wheat and corn light and dark dark and light

4. Conclusions

We have introduced the first Hungarian language end-to-end speech recognition results on high-quality spontaneous speech. The subset of BEA database used for training and evaluation is available for research purposes. We showed

that ASR error metrics are significantly higher for spontaneous speech than for the planned one even in the case of identical speakers in the same environment. The results suggest that spontaneous speech needs special care in modelling and/or more data in order to cope with the high variability of spontaneously articulated or fast speech. The validity of the findings is limited though by the fact that the planned part of the database is based on the same text for each speaker. As for future work, we are going to re-sample the planned part of the corpora to decrease the mismatch of planned/spontaneous evaluation. Also, we plan to add a language model to the model/decoder since it obviously could reduce the word error rates according to the results of subjective evaluation. All in all, the spontaneous results can serve as baseline for future improvements.

References

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, *5(2)*, 179–190.

Ginsburg, B., Castonguay, P., Hrinchuk, O., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Nguyen, H., Zhang, Y., & Cohen, J. M. (2020). Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. *arXiv*, *1905.11286*.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369–376).

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T. E., Beke, A., Neuberger, T., & Nikléczi, P. (2012). BEA: Beszélt nyelvi adatbázis. In M. Gósy (Ed.), *Beszéd, adatbázis, kutatások* (pp. 9–25). Budapest: Akadémiai Kiadó.

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., & Zhang, Y. (2019). QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions. *arXiv*, *1910.10261*.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., & Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *arXiv*, *1904.03288*.

Mihajlik, P. (2020). How does an AI recognize speech?–about end-to-end deep neural network based speech recognition. In *Speech Research conference* (pp. 68–70). volume 14.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). doi:10.1109/ICASSP.2015.7178964.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. *Interspeech*, *2019*. URL: http://dx.doi.org/10.21437/Interspeech.2019-2680.

Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., & Collobert, R. (2019). Wav2letter++: A fast open-source speech recognition system. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2019*. URL: http://dx.doi.org/10.1109/ICASSP.2019.8683535. doi:10.1109/icassp.2019.8683535.