Recognition of the Electrolaryngeal Speech: Comparison Between Human and Machine

Petr Stanislav(✉), Josef V. Psutka, and Josef Psutka

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia, Pilsen, Czech Republic
{pstanisl,psutka_j,psutka}@kky.zcu.cz
http://www.kky.zcu.cz/en

Abstract. Automatic recognition of an electrolaryngeal speech is usually a hard task due to the fact that all phonemes tend to be voiced. However, using a strong language model (LM) for continuous speech recognition task, we can achieve satisfactory recognition accuracy. On the other hand, the recognition of isolated words or phrase sentences containing only several words poses a problem, as in this case, the LM does not have a chance to properly support the recognition. At the same time, the recognition of short phrases has a great practical potential. In this paper, we would like to discuss poor performance of the electrolaryngeal speech automatic speech recognition (ASR), especially for isolated words. By comparing the results achieved by humans and the ASR system, we will attempt to show that even humans are unable to distinguish the identity of the word, differing only in voicing, always correctly. We describe three experiments: the one represents blind recognition, i.e., the ability to correctly recognize an isolated word selected from a vocabulary of more than a million words. The second experiment shows results achieved when there is some additional knowledge about the task, specifically, when the recognition vocabulary is reduced only to words that actually are included in the test. And the third test evaluates the ability to distinguish two similar words (differing only in voicing) for both the human and the ASR system.

Keywords: Electrolaryngeal speech · ASR · Listening tests

1 Introduction

Surgical removal of the vocal cords during total laryngectomy has a major impact on the quality of life of a person. After such surgery people are unable to produce a voiced sound, their speech is called alaryngeal speech. There are several methods of restoring the speech after total laryngectomy (TL). Esophageal speech belongs to the most common methods used for speech restoration. The idea is based on releasing gases from esophagus instead of lungs. But this method has a low acquisition rate (only ∼6% of the patients are able to learn it [6]). Another

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 509–517, 2017.
DOI: 10.1007/978-3-319-64206-2_57

method uses a tracheoesophageal prosthesis that connects the larynx with pharynx. The air passing into the pharynx causes the required vibrations and the voiced sound can be created [5].

Another option to produce the excitation necessary for proper voicing is using an external device - the electrolarynx (EL). EL is a battery-powered device that mechanically generates sound source signals which are conducted into the oral cavity from either the soft parts of the neck or from the lower jaw [9]. In most cases, a monotone pitch is generated and the produced speech sounds very mechanical. Moreover, in order to make the resulting speech sufficiently audible, the EL needs to generate rather strong vibrations that are also emitted outside of the speaker's body and could be perceived as noise by other people. Unfortunately, neither of these methods of rehabilitation can guarantee the natural sound of the resulting voice.

The great disadvantage of using EL is that all pronounced phonemes tend to be voiced. That's why the automatic recognition of the EL speech is usually a very hard task. Nevertheless, we can achieve good results but only in cases where we can use a strong language model (LM) during recognition. Unfortunately, the LM can fully unfold its strength only when recognizing longer sentence segments where it can make use of the word context. But the problem is that speakers with TL almost do not communicate. And if they do, they do so only in short phrases or instructions. This is mainly due to the unnatural sound of their voice which they are often ashamed of. The second reason is that shorter sentences are less exhausting for EL speakers. In this case, the benefit of the LM seems to be negligible (it is hard to determine the identity of words without the necessary context). However, such task has a great practical potential.

In this paper, we would like to compare the ability of humans and machines (equipped with ASR system) to distinguish the identity of acoustically very similar words and word bigrams. For this purpose, we created two listening tests. The first one contains 320 isolated words and the second one contains 333 word bigrams. Both tests are described in detail (including the results) in Sect. 3. In Sect. 4, we present the process of the ASR system training and ASR results on test sentences. And finally in Sect. 5, we offer the comparison between human and machine recognition results.

2 Database Description

Although there are various ways to rehabilitate the voice [1], none of them is perfect. One possible way of improving the patient's ability to communicate is to use ASR and speech synthesis (TTS) technology [4] in some means of communication, e.g., for phone calls, but the biggest issue is the lack of speech data for training the ASR acoustic models. In comparison with healthy speakers, speech recording is naturally significantly more exhausting for people with TL using the electrolarynx and it is therefore complicated to obtain the relevant amount of data.

For the purpose of our research, we recorded a speech corpus which contains 5589 sentences and 320 isolated words from one speaker. It consists of about

14 h of annotated speech. The speaker is an older woman who underwent TL more than 10 years ago and uses electrolarynx (EL) on a daily basis. All the sentences and words are in the Czech language. The database of text prompts from which the sentences were selected was obtained in an electronic form from the web pages of Czech newspaper publishers [8].

The corpus consists of two parts. The major part contains 40 phonetically rich and 5036 phonetically balanced sentences. The other part contains 419 sentences and 320 isolated words. Separation of the corpus into two parts was necessary due to the large time span between the recordings. During that period, the speaker changed the type of the electrolarynx, which plays an important role in the quality of the speech, and also the recording technology was upgraded.

Part of the corpus with isolated words contains 160 pairs of specifically selected words that have different meaning but they acoustically differ only in voicing. An example is a pair of Czech words "kosa[1]" - "koza[2]". For each word, we recorded at least one sentence containing the word.

All the speech utterances are sampled at 48 kHz and 16 bit amplitude resolution and resampled to a sampling frequency of 16 kHz for the speech recognition task.

3 Listening Tests

For our human vs. machine comparison, it is crucial to obtain relevant results from as large group of listeners as possible. For this purpose, we created two listening tests. The first one contained recordings of isolated words from the corpus mentioned above, the second one the combination of word pairs that differ only in the voicing of exactly one phoneme. The task was to correctly recognize the identity of recorded words. In both cases, participants were asked to choose one of the prepared answers through a web-based interface. Subjects were allowed to replay the speech recording as many times as they wanted. The participation in the test was voluntary and the subjects were predominantly fellow researchers from our lab.

The listening tests have been designed to verify the human ability to distinguish two words having various meanings but differing only in the voicing of a single phoneme.

Such task is rather easy when listening to "natural" speech but in the case of electrolaryngeal speech, the situation is quite different. The reader should bear in mind that the EL produces the excitation signal constantly and therefore even the originally voiceless phonemes should theoretically become voiced. The form of isolated words and word bigrams was chosen intentionally because it is much harder to determine the meaning – and thus correctly recognize the word identity – when there is no context.

[1] Scythe.

[2] Goat.

3.1 Isolated Words

In the test, participants were asked to listen to 320 recordings of isolated words and select one of the prepared options which were in the form: (a) *word A (e.g. kosa)*, (b) *word B (e.g. koza)*, (c) *cannot decide*. The first two options represent a combination of the word that was really uttered in the recording and a complementary word differing in a single phoneme voicing. These options were in all cases presented in the alphabetical order. If the participants had no clue about the identity of the replayed word, they were instructed to choose the third option (cannot decide). The samples were presented to each listener in random order. The test was completed by 19 subjects.

The output of the test is a table with percentages of responses for each option. Table 1 shows an excerpt from the results. The percentages of the correct answers are highlighted by boldface. The first example represents situations where participants were unable to make a clear decision about the word identity. The second example represents a pair of complementary words where all the participants selected in all cases the same option, regardless of the word that was actually uttered. It could be interpreted as that the acoustic form of the Czech word "kosa" is exactly the same as the word "koza" if the speaker uses the EL. The last example shows a pair of words for which the participants were able to identify the word correctly in almost all cases. The overall accuracy of answers was $Acc_w^{human} = 70,47\%$ and it was counted as follows

$$Acc_w^{human} = \frac{1}{n}\sum_{i=1}^{n} f_i * 100, \tag{1}$$

where $n = 320$ and f_i is equal to a relative frequency of the correct answer to the question i in the listening test.

Table 1. Sample results of the listening test with the isolated words with percentage of selected options. The percentages of the correct answers are highlighted by boldface.

Word	Option 1	Option 2	Option 3
borce	**57.90**	36.84	5.26
porce	21.05	**52.63**	26.32
kosa	**0.00**	100.00	0.00
koza	0.00	**100.00**	0.00
přibít	**94.74**	5.26	0.00
připít	10.52	**89.48**	0.00

3.2 Word Bigrams

In the test, participants were asked to listen 333 recordings of the word bigrams (two consecutively uttered words) and select one of the prepared options that have the forms (a) *word A + word A (e.g. kosa + kosa)*, (b) *word A + word B (e.g. kosa + koza)*, (c) *word B + word A (e.g. koza + kosa)*, (d) *word B + word B (e.g. koza + koza)*. It can be seen that this represents all the combinations of the words from the bigram and also that the bigrams were always assembled using the words from the complementary pairs differing only in voicing of a single phoneme. An audio file containing the bigram is a concatenation of two recordings of the isolated words with a short pause between them. Higher number of questions in the test is due to the fact that for some words there exists more than one possible combination with another word. In an effort to shorten the time of the already long test, we generated only the combinations of words corresponding to the second and third option in the test; the participants were not informed about this fact. However, the listening test was completed only by 12 subjects.

The output of the test is again a table with percentages of responses for each option. Table 2 shows an excerpt from the results. As in the previous case, the percentages of the correct answers are highlighted by boldface. It depicts the results for the same words that were shown for isolated word scenario in Table 1. Although now the test deals with word bigrams, the results from both tests naturally exhibit a correlation. For the first combination of words *"borci"* and *"porci"*, the participants were unable to convincingly decide about the content of the recording, just as they were not able to clearly identify those words in the first listening test. In the second example, all the participants except one voted for the combination of words *"koza + koza"* even though the bigrams contained both words from this complementary pair. The last example is an illustration of the situation where the participants were able to clearly distinguish the words. The overall accuracy of correct answers is $Acc_p^{human} = 66,24\%$ and it was computed using Eq. 1.

Table 2. Sample results of the listening test with the word bigrams with percentage of selected options. The percentages of the correct answers are highlighted by boldface.

Word bigram	Option 1	Option 2	Option 3	Option 4
borce + porce	16.67	**50.00**	0.00	33.33
porce + borce	8.33	0.00	**66.67**	25.00
kosa + koza	0.00	**8.33**	0.00	91.67
koza + kosa	0.00	0.00	**0.00**	100.00
přibít + připít	0.00	**100.00**	0.00	0.00
připít + přibít	0.00	0.00	**100.00**	0.00

4 Acoustic Modeling and Recognition Results

We followed the typical Kaldi [7] training recipe S5 [3] for a deep neural network (DNN) acoustic model (AM) training. In all our experiments we used features based on PLP parameterization (19 band pass filters, 12 cepstral coefficients with delta and delta-delta features with CMN). The first step of the Kaldi training recipe is a monophone acoustic model. A monophone AM is trained from the flat start using the PLPs features (static + delta + delta delta) and is treated as a special case of a context-dependent system, without any left or right context. Secondly, we train the triphone GMM AM. As the number of triphones is too large, we used decision trees to tie their states. The questions can be based on linguistic knowledge, but in our case, the questions were generated fully automatically (based on a tree-clustering of the phones). Due to a lack of acoustic training data, we performed several experiments to determine the best WER according to the number of clustered states. The best results were obtained for 2048 states.

We also applied linear discriminant analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) over a central frame spliced across 3 frames. LDA+MLLT project the concatenated frames into 40 dimensions space. Despite the fact that we have only one speaker, we used feature space Maximum Likelihood Linear Regression (fMLLR) [2] because the recordings were taken in two separate series; the individual series differ not only in the recording equipment but also in speaker's EL.

The 40-dimensional features from GMM are spliced across 5 frames of context and used as input to the DNN. The resulting dimension of the input feature vector of the DNN is therefore 440. In this framework, we used the standard 6 layers topology (5 hidden layers, each with 2048 neurons). The output layer was a softmax layer with dimension equal to the number of clustered context-dependent states (the number of states was one of the task's parameters that were optimized).

The S5 recipe supports layer-wise restricted Boltzmann machine (RBM) pre-training, stochastic gradient descent (SGD) training and sequence-discriminative training, using lattice framework, optimizing state-level minimum Bayes risk (sMBR) criterion. The whole DNN training is running on a single GPU using CUDA. The training dataset consists of a total of 5000 sentences from both parts of the corpus.

For all recognition experiments, we used our own RT decoder. This in-house LVCSR system is optimized for low latency in the real-time operation with very large vocabularies. To enable recognition with a vocabulary containing more than one million words in real-time, we accelerated the decoding using parallel approach (Viterbi search on CPU and DNN segments scores on GPU). We used trigram back-off language models (LM) with mixed-case vocabularies with more than 1.2M words. Our text corpus containing the data from newspapers (520 million tokens), web news (350 million tokens), subtitles (200 million tokens) and transcriptions of some TV programs (175 million tokens) (details can be found in [10]). The overall accuracy on the test dataset containing 580 of randomly selected sentences is $Acc_w^{machine} = 86.10\%$ and it was computed as follows

$$Acc_w^{machine} = \frac{N - S - D - I}{N} * 100, \tag{2}$$

where N is the number of words, S is the number of substitutions, D is the number of deletations and I is the number of insertions.

5 Experiments and Evaluation

As mentioned above, the goal is to compare the ability of the human and the machine to distinguish the identity of words, differing only in voicing. In this section, we describe three experiments designed for such comparison and present the evaluation of the results.

From the listening tests described above, we obtained two sets of results. The first one represents the ability to determine the identity of the word pronounced in isolation, and the second one the ability to distinguish two similar words pronounced together. We created three ASR experiments that should emulate the same scenarios, using the Kaldi model described in Sect. 4.

The first experiment with isolated words and the zerogram LM[3] containing more than a million words should correspond to what we called a blind test. In the real situation, the listener does not know the word in advance, which is the reason why such a large vocabulary is used. We named the experiment *"onemil"*. However, in the actual listening test, the participants knew the list of included words and therefore had some advantage over the *"onemil"* machine setting. To compensate for this, we reduced the size of ASR vocabulary and included only the words that actually occurred in the test. We named it *"reduced"*. Both experiments are compared with the first listening test. Note that another possible solution would be to create a special vocabulary with just the specific complementary word pair. But since the listening test contains three possible answers (including the "cannot decide" option), and the ASR system with a two-word vocabulary would always select one of the words from the pair, the results would not correspond to the listening test. The last experiment corresponds to the second listening test. To obtain comparable results, we generated special LM for each word bigram. The LM contains only all four combinations of the words, same as the listening test. The output is the best-matched combination. We named the experiment *"bigrams"*.

The results from the recognizer are not directly comparable with those of listening tests. For resolving the issue we rated correct hypothesis with 1, otherwise 0, and computed the average. The overall results are presented in Table 3.

Table 3. Average accuracy results in percent for human and machine.

	onemil	reduced	bigrams
human	70.47	70.47	66.24
machine	61.76	69.91	54.82

[3] Zerogram LM is the setting where all words from a fixed vocabulary have the same probability.

The results show that the presented tasks are challenging even for humans. In the case of the *"onemil"* experiment, the performance of the ASR is further hurt by the enormous perplexity of the language model (note that for zerogram LM the perplexity is directly equal to the vocabulary size). Reducing the size of the vocabulary in the *"reduced"* experiment lifted the ASR accuracy almost to the human level; let us stress out that even this scenario is unfair to ASR as the humans are effectively facing only the perplexity of 3.

Interesting are the results of the experiment with word bigrams. At first glance, it may seem easier because the task is to select well-defined combinations of words. However, words are acoustically very similar and this makes it very difficult to tell them apart. In many cases, the difference of rating of the ASR hypotheses is very small. It indicates similarity between the incriminated models of phonemes. It occurs mostly in bigrams where humans were not able to conclusively decide.

6 Conclusion

In this paper, we compared the ability of humans and machines to distinguish the identity of acoustically similar words.

One important conclusion is that the recognition of isolated words pronounced using the electrolarynx is a hard task even for humans and nowadays ASR systems are not better. In the ideal situation, we are able to achieve competitive results (i.e. 86.10% accuracy) if there is sufficient context. However, with a smaller (or non-existent) context, the recognition quality dramatically decreases (both for humans and machines). An illustrative example is the performance drop in an isolated word scenario. On the other hand, after the experience with corpus recording, we should suppose that the communication using short phrases (and the resulting limited word context) is a realistic scenario when dealing with EL speakers – most of the recorded sentences were split into multiple parts during the post-processing since the speakers needed to make pauses for recovery in the middle of long sentences.

Some relatively significant differences between human and machine performance could be caused by the aforementioned lower effective perplexity in the human task, stemming from the listening tests design. Some participants may also unravel the fact that a test always contains only a subset of all possible answer variants. Also, the human participants could listen to the recording several times and then select an answer. However, we must honestly say that the ASR system as prepared for our experiments can hardly benefit from repeated processing of the same audio file.

All those drawbacks could be solved by a more sophisticated listening test, but the problem is the difficulty of even the existing test. The slightly more challenging listening test with word bigrams was completed only by two-thirds of the participants of the first test.

At the beginning of the research, we assumed all phonemes are voiced in EL speech. Experiments have shown that this is not always true. The context of the phoneme plays an important role in voicing. The human and the machine were able, in many cases, to distinguish voiced and unvoiced phonemes in the EL speech. The processing of these phonemes will be one of the subjects of further research.

Achieved results with an ASR system lead to the conclusion that if some preprocessing is done or special models are created, EL users can have access to ASR technologies.

Acknowledgements. The work has been supported by the grant of the University of West Bohemia, project No. SGS-2016-039 and by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506.

References

1. Brown, D.H., Hilgers, F.J., Irish, J.C., Balm, A.J.: Postlaryngectomy voice rehabilitation: state of the art at the Millennium. World J. Surg. **27**(7), 824–831 (2003)
2. Fuchs, A.K., Morales-Cordovilla, J.A., Hagmller, M.: ASR for electro-laryngeal speech. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 234–238, December 2013
3. Github Kaldi: https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5
4. Jůzová, M., Romportl, J., Tihelka, D.: Speech Corpus Preparation for Voice Banking of Laryngectomised Patients, pp. 282–290. Springer, Cham (2015)
5. Kramp, B., Dommerich, S.: Tracheostomy cannulas and voice prosthesis. GMS Curr. Top. Otorhinolaryngol. Head Neck Surg. **8**, Doc05 (2009)
6. Liu, H., Ng, M.L.: Electrolarynx in voice rehabilitation. Auris Nasus Larynx **34**(3), 327–332 (2007)
7. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. No. EPFL-CONF-192584. IEEE Signal Processing Society (2011)
8. Radová, V., Psutka, J.: UWB-S01 corpus: a Czech read-speech corpus. In: Proceedings of the 6th International Conference on Spoken Language Processing, ICSLP 2000, pp. 732–735 (2000)
9. Stanislav, P., Psutka, J.V.: Influence of different phoneme mappings on the recognition accuracy of electrolaryngeal speech. In: Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems, SIGMAP (ICETE 2012), vol. 1, pp. 204–207 (2012)
10. Švec, J., Hoidekr, J., Soutner, D., Vavruška, J.: Web Text Data Mining for Building Large Scale Language Modelling Corpus, pp. 356–363. Springer, Heidelberg (2011)