Influence of different phoneme mappings on the recognition accuracy of electrolaryngeal speech

Petr Stanislav[1]

1 INTRODUCTION

A malignant disease of vocal folds does not occur as often as for example breast cancer or lung cancer. However, If treatment is not successful, the consequences of this illness could be very serious. In extreme cases, the total laryngectomy is not able to speak using his own vocal folds.

There are several methods of restoring the speech for total laryngectomees. One of the this methods is to produce the necessary excitations using an external device - the electrolarynx.

2 TOTAL LARYNGECTOMEES

The total larygectomy is a surgery during which the vocal folds affected by cancer are removed. The healthy vocal folds excite a stream of air from the lungs and then the excited stream is modulated in the nasal and oral cavity and speak through the mouth. However, in case of total laryngectomees, there is no connection between the larynx and pharynx. Therefore the flow of air does not flow form the lungs to the mouth. Therefore the speech could not be produced in the same way as in the case of nonlarygectomees Nakamura (2010).

One way of replacing removed vocal folds is to use an electromechanical device called electrolarynx. The patient attaches the electrolarynx either to the soft parts of the neck or to the lower jaw and the vibrating plate substitutes the missing vocal fold vibrations. This method is easily manageable. Yet this method still has some flaws, for example the monotonous mechanical voice of a speaker.

3 INFLUENCE OF THE PHONEME MAPPING

A laryngectomee uses electrolarynx is not able to speak when the device is off. And since the vibrating plate provides constant excitation, it is not possible for him to produce unvoiced phonemes. To verifying this assumption were speech data obtained from two female speakers, one was a person with healthy voice and one total laryngectomee and phoneme mapping technique was used.

Two different approaches were tested and compared together. In the first one the basic speech unit was monophones in contrast to triphones in the second one. The special systems using phonemes mapping were built for testing of speech recognition. In this case no difference between results given by system without mapping and phonemes mapping system should be detected in case of laryngectomee. In specific case the accuracy of recognition could be improved due to reduction of the system perplexity. Conversely, in case of nonlarygectomees the reduction of the phonetic set could lead to reducing the accuracy.

Obtained recognition accuracy for monophone models without/with mapping with zerogram based language model is written in Table 1 and Table 2. From these tables it could be seen

[1] student navazujícího doktorského studijního programu Aplikované vědy a informatika, obor Mechanika, specializace Aplikovaná mechanika, e-mail: pstanisl@kky.zcu.cz

that every change of phonetic set causes reducing of speech recognition accuracy for nonlaryngectomee (last column). However, for total laryngectomees it is not possible to confirm this assumption clearly. From computed results it is possible to obtain information about accuracy, thus about decreasing of accuracy due to replacing unvoiced monophones/triphones by voiced one. The same character of result was obtained from phoneme mapping ’t’, ’ť’ → ’d’, ’ď’ and ’f’ → ’v’. Conversely, if ’k’ was replaced by ’g’ then the higher speech recognition accuracy was obtained than for baseline model. From replacing ’s’,’š’ → ’z’,’ž’ the obtained results were not clear.

Acoustic model	L. no map.[%]	NL no map. [%]	L. w. map. [%]	NL w. map. [%]
Baseline	83.05	91.35	84.92	86.47
’f’ → ’v’	83.05	89.96	84.51	87.42
’k’ → ’g’	83.10	90.58	85.50	86.36
’s, ’š’ → ’z’, ’ž’	83.71	88.77	84.75	84.81
’t’, ’ť’ → ’d’, ’ď’	82.47	90.05	84.38	86.38
All voiced	82.78	86.58	84.34	83.77

Tabulka 1: Accuracy of the ASR system with monophone acoustic models and zerogram based language model for laryngectomee speaker and nonlaryngectomee speaker.

Acoustic model	L. no map.[%]	NL no map. [%]	L. w. map. [%]	NL w. map. [%]
Baseline	82.60	92.66	87.65	95.80
’f’ → ’v’	82.23	92.41	87.51	95.46
’k’ → ’g’	83.30	92.57	88.38	95.55
’s’, ’š’ → ’z’, ’ž’	83.28	92.28	88.31	95.07
’t’, ’ť’ → ’d’, ’ď’	82.13	92.28	87.60	95.39
All voiced	82.18	91.03	86.97	94.53

Tabulka 2: Accuracy of the ASR system with triphone acoustic models and zerogram based language model for laryngectomee and nonlaryngectomee.

4 CONCLUSION

We have focused on the problem, with voiced and unvoiced phonemes. The test results for both monophone- and triphone-based acoustic models showed that the substitution of all unvoiced phonemes for voiced ones decreased recognition accuracy for both language models. But on the other hand there were phoneme substitutions that increases the accuracy. The interesting issue is how can for instance substitution ’s’, ’š’ → ’z’, ’ž’ give better recognition results in tests with monopohone-based than in the triphone-based acoustic models in comparison to baseline acoustic models. This can be due to a more complex phonetic structure in triphone-based acoustic model that can represent small differences between phonemes in different surroundings even if there are pronounced as voiced sound.

Literatura

Nakamura, K, 2010. *Doctoral Thesis: Speaking Aid System Using Statistical Voice Conversion for Electrolaryngeal Speech*. Japan.