Influence of Different Phoneme Mappings on the Recognition Accuracy of Electrolaryngeal Speech

Petr Stanislav and Josef V. Psutka
Department of Cybernetics, University of West Bohemia, Univerzitní 8, 306 14 Pilsen, Czech Republic

Keywords: Automatic Speech Recognition, Laryngectomees, Electrolaryngeal Speech, Phoneme Mapping.

Abstract: This paper presents the initial steps towards building speech recognition system that is able to efficiently process electrolaryngeal substitute speech produced by laryngectomees. Speakers after total laryngectomy are characterized by restricted aero-acoustic properties in comparison with normal speakers and their speech is therefore far less intelligible. We suggested and tested several approaches to acoustic modeling within the ASR system that would be able to cope with this lower intelligibility. Comparative experiments were also performed on the healthy speakers. We tried several mappings that unify unvoiced phonemes with their voiced counterparts in the acoustic modeling process both on monophone and triphone level. Systems using zerogram and trigram language models were evaluated and compared in order to increase the credibility of the results.

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 (which includes the removal of vocal folds) is performed. Therefore the person who undergoes this surgery is not able to speak in a standard way.

There are several methods of restoring the speech for total laryngectomees. Esophageal speech belongs to the most common methods used for speech restoration. The idea is based on releasing gases from esophagus instead of lungs. Another method uses a tracheoesophageal prosthesis that connects the larynx with pharynx. The air passing into the pharynx causes the required vibrations and the utterance can be created. Another option how to produce the necessary excitations is using an external device - the electrolarynx.

This paper describes the influence of the Czech phoneme/triphone mapping on the accuracy of the speech recognition results of a total laryngectomee. The obtained results are compared with accuracy of the speech recognition of a healthy person using the same phoneme/triphone mapping. In Section 2, the difference between speech production of the nonlaryngectomme and laryngectomee speaker is described. Section 3 explains the process of the acoustic model creation and principle of the phoneme mapping. Section 4 presents obtained results and the Section 5 concludes the paper.

2 TOTAL LARYNGECTOMEES

The total larygectomy is a surgery during which the vocal folds affected by cancer are removed. The differences between the healthy speaker and speaker without vocal folds are shown in Figure 1 and Figure 2, respectively. 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. The modulated stream comes out from the mouth as speech. However, in case of total laryngectomees, there is no connection between the larynx and the oral cavity. Therefore the flow of air does not flow form the lungs to the mouth, but to the tracheostoma that is used for breathing. 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 electrolarynx is useful for total laryngectomees who have not obtained any tracheoesophageal prosthesis and have not been able to learn an esophageal speech.

The basic part of this device is a battery powered electric motor that excites a vibration plate. The pa-

Stanislav P. and V. Psutka J..
Influence of Different Phoneme Mappings on the Recognition Accuracy of Electrolaryngeal Speech.
DOI: 10.5220/0004129502040207
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications and Wireless Information Networks and Systems* (SIGMAP-2012), pages 204-207
ISBN: 978-989-8565-25-9

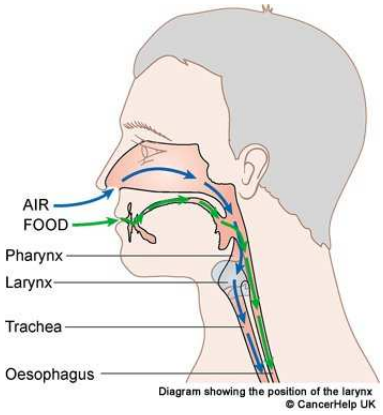

Figure 1: Scheme of ingestion and breathing for nonlaryngectomees.

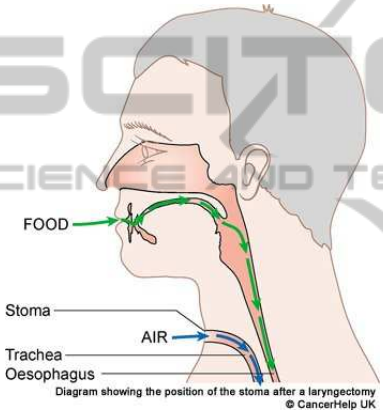

Figure 2: Scheme of ingestion and breathing for total laryngectomees.

tient 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 as it is shown in Figure 3. This method is easily manageable. After a very short time, the speaker is able to produce continuous speech. Yet this method still has some flaws, for example the monotonous mechanical voice of a speaker or the poor speech intelligibility in a noisy environment due to a constant volume level. Moreover, the electrolarynx creates continuous sound that might be irritating and one hand of the speaker is always occupied holding the device while speaking.

3 ACOUSTIC MODELS

The aim of this paper was to verify the assumption that the laryngectomees who use an electrolarynx produce voiced phonemes only (because the electrolarynx provides continuous excitation). Source data were obtained from two female speakers, one was

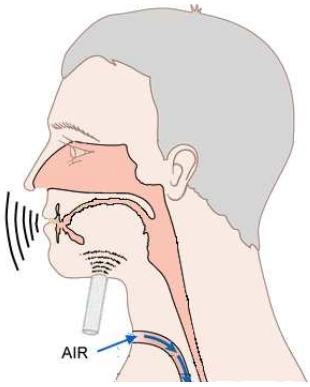

Figure 3: Usage of electrolarynx.

a person with healthy voice and one total laryngectomee. The female who underwent total laryngectomy did so already 10 years ago and was able to speak with electrolarynx in long utterances.

Both women read 5000 sentences which amounts to more than 10 hours of speech from each speaker. The source texts were selected from the database that was created from the web pages of Czech newspaper publishers(Radová and Psutka, 2000). Special consideration was given to the sentence selection as we wanted to have a representative distribution of the more frequent triphone sequences (reflecting their relative occurrence in natural speech). The corpus was recorded in the office where only the speaker was present.

The digitization of the analogue signal was provided at 44.1 kHz sample rate and 16-bit resolution format by the special DPA miniature omnidirectional microphone. The front-end worked with MFCC parameterization with 26 filters and 12 MFCC cepstral coefficients plus energy with both delta and delta-delta sub-features (see (Psutka and et al., 2007) for methodology). Therefore one feature vector contains 39 coefficients. Feature vectors are computed each 10 milliseconds (100 frames per second).

If a laryngectomee uses electrolarynx for speech production, he 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. This assumption was verified by recording isolated Czech words that differ only in voicing - 'koza' and 'kosa'('goat' and 'scythe' in English), where 'z' is voiced and 's' is unvoiced. There was no audible difference between both utterances. Comparing the acoustic properties also did not reveal any significant difference between analyzed words. Therefore selected unvoiced triphones/phonemes were replaced by corresponding voiced ones (see Table 1) in the acoustic modeling process.

Table 1: Corresponding pairs of phonemes.

Unvoiced phoneme	Voiced phoneme
f	v
k	g
s	z
š	ž
t	d
ť	ď

4 EXPERIMENTS

Two different approaches were tested and compared together (for both training corpuses). In the first one the basic speech unit was monophone in contrast to triphone in the second one. In all our experiments the individual basic speech unit was represented by a three-state HMM with a continuous output probability density function assigned to each state. As the number of the Czech triphones is large, phonetic decision trees were used to tie the states of the triphones. Several experiments were performed to determine the best recognition results according to the number of clustered states and also to the number of mixtures. The prime Gaussians triphone/monophone acoustic model trained with the Maximum Likelihood (ML) criterion was made with HTK-Toolkit v.3.4.

The special systems using phonemes mapping were built for testing of speech recognition. The main idea of the experiment is based on the vocalization of all produced phonemes. In this case no difference between results given by system without mapping and phonemes mapping system should be detected. In specific case the accuracy of recognition could even be improved due to reduction of the system perplexity. The system does not use the full phonetic set.

Conversely, in case of nonlarygectomees the reduction of the phonetic set could lead to reducing the accuracy. Remember that the source data were chosen with an emphasis of inclusion of all Czech triphone/monophone in corresponding representation.

The test set consists of 500 sentences for both training corpuses (nonlarygectomees and laryngectomee speech). This portion of sentences (10% of the original training set) contains approximately 1 hour of speech for each speaker. In all recognition experiments, a language model based on zerogram as well as a trigram-based one were applied in order to judge a quality of developed acoustic models. The perplexity of the zerogram language model was 2885 (in other words, the recognition lexicon contained 2885 words) and there were no OOV words. The trigram languages models were trained by SRI Language Modeling Toolkit (Stolcke, 2002) using modified Kneser-Ney smoothing that proved to be efficient in our previous language modeling experiments (Pražák et al., 2008). We have collected large 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). The model contained the most frequent 360K words with OOV amounting to 3.8%. The perplexity of the recognition task was 3380.

The verification of the assumption was realized by an acoustic models using triphone/monophone for the speech recognition. All models are created for both speakers. Firstly the baseline acoustic model without mapping was created. Then the model that maps only voiceless triphones/monophones. Due to identification of the influence of each phoneme on system accuracy four more models were built.

- acoustic model with mapping ’f’ on ’v’;
- acoustic model with mapping ’k’ on *g*;
- acoustic model with mapping ’s’,’ š’ on ’z’, ’ž’;
- acoustic model with mapping ’t’, ’ť’ on ’d’, ’ď’;

For verification of our assumptions 24 acoustic models were created (6 monophone model and 6 triphone models for each speaker). Obtained recognition accuracy is given in Table 2 for monophone model with zerogram based language model in case of and Table 3 for monophone model with trigram language model with 360K words lexicon.

From these tables it could be seen that every change of phonetic set causes reducing of speech recognition accuracy for nonlaryngectomee. 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’ $\rightarrow$ ’d’, ’ď’ and ’f’ $\rightarrow$ ’v’. Conversely, if ’k’ was replaced by ’g’ then the higher speech recognition accuracy was obtained than for baseline model. From replacing ’s’,’š’ $\rightarrow$ ’z’,’ž’ the obtained results were not clear. Therefore the further work will be focused on solution of this problem.

5 CONCLUSIONS

We have presented our initial investigations into the challenging problem of transcribing electrolaryngeal substitute speech of total laryngectomees. We have

Table 2: Accuracy of the ASR system with monophone acoustic models and zerogram based language model for laryngectomee speaker and nonlaryngectomee speaker.

Acoustic model	Laryng. [%]	Nonlaryng. [%]
Baseline	83.05	91.35
'f' → 'v'	83.05	89.96
'k' → 'g'	83.10	90.58
's, 'š' → 'z', 'ž'	83.71	88.77
't', 'ť' → 'd', 'ď'	82.47	90.05
All voiced	82.78	86.58

Table 3: Accuracy of the ASR system with monophone acoustic models and trigram based language model containing 360k words for laryngectomee and nonlaryngectomeer.

Acoustic model	Laryng. [%]	Nonlaryng. [%]
Baseline	84.92	87.47
'f' → 'v'	84.51	87.42
'k' → 'g'	85.50	86.36
's', 'š' → 'z', 'ž'	84.75	84.81
't', 'ť' → 'd', 'ď'	84.38	86.38
All voiced	84.34	83.77

Table 4: Accuracy of the ASR system with triphone acoustic model s and zerogram based language model for laryngectomee and nonlaryngectomee.

Acoustic model	Laryng. [%]	Nonlaryng. [%]
Baseline	82.60	92.66
'f' → 'v'	82.23	92.41
'k' → 'g'	83.30	92.57
's', 'š' → 'z', 'ž'	83.28	92.28
't', 'ť' → 'd', 'ď'	82.13	92.28
All voiced	82.18	91.03

Table 5: Accuracy of the ASR system with triphone acoustic models and trigram based language model containing 360k words for, laryngectomee and nonlaryngectomeer.

Acoustic model	Laryng. [%]	Nonlaryng. [%]
Baseline	87.65	95.80
'f' → 'v'	87.51	95.46
'k' → 'g'	88.38	95.55
's', 'š' → 'z', 'ž'	88.31	95.07
't', 'ť' → 'd', 'ď'	87.60	95.39
All voiced	86.97	94.53

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 (e.g. 'k' → 'g') that increased 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. We would like to investigate such interesting issues in our future work.

ACKNOWLEDGEMENTS

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 and by the grant of the University of West Bohemia and project No. SGS-2010-054.

REFERENCES

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

Pražák, A., Ircing, P., Švec, J., and Psutka, J. V. (2008). Efficient combination of n-gram language models and recognition grammars in real-time lvcsr decoder. In *9th International Conference on Signal Processing Proceedings*, pages 587–591, Peking, China. IEEE.

Psutka, J. V. and et al. (2007). Searching for a robust mfcc-based parameterization for asr application. *SIGMAP 2007: Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 196–199.

Radová, V. and Psutka, J. (2000). UWB-S01 corpus: A czech read-speech corpus. *Proceedings of the 6th International Conference on Spoken Language Processing*.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. *International Conference on Spoken Language Processing.*