published in: Text, Speech and Dialogue (TSD 2004), Proceedings
LNAI 3206, Springer, Berlin, Heidelberg, 2004, pp. 331–338

Automatic Recognition and Evaluation of Tracheoesophageal Speech *

Tino Haderlein[1], Stefan Steidl[1], Elmar Nöth[1], Frank Rosanowski[2], and Maria Schuster[2]

[1] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] Abt. für Phoniatrie und Pädaudiologie des Universitätsklinikums Erlangen Bohlenplatz 21, 91054 Erlangen, Germany

Abstract. Tracheoesophageal (TE) speech is a possibility to restore the ability to speak after laryngectomy, i.e. the removal of the larynx. TE speech often shows low audibility and intelligibility which also makes it a challenge to automatic speech recognition. We improved the recognition results by adapting a speech recognizer trained on normal, non-pathologic voices to single TE speakers by unsupervised HMM interpolation.
In speech rehabilitation the patient's voice quality has to be evaluated. As no objective classification means exists until now and an automation of this procedure is desirable we performed initial experiments for automatic evaluation of the intelligibility. We compared scoring results for TE speech from five experienced raters with the word accuracy from different types of speech recognizers. Correlation coefficients of about -0.8 are promising for future work.

1 Introduction

The results of a speech recognition task depend on the quality of the input signal. The term "quality" is in this context mostly used in the frame of influences by the transmission channel or background noise, but the speaker's voice can be the source of recognition problems as well. This paper focuses on the recognition of a special kind of pathologic voices, i.e. tracheoesophageal (TE) voices. In tracheoesophageal speech, the upper esophagus, the pharyngo-esophageal (PE) segment, serves as a sound generator (see Fig. 1). The air stream from the lungs is deviated into the esophagus during expiration via a shunt between the trachea and the esophagus. Tissue vibrations of the PE segment modulate the streaming air and generate a substitute voice signal. In comparison to normal voices the

* This work was partly funded by the EU in the project PF-STAR under grant IST-2001-37599. The responsibility for the contents of this study lies with the authors.

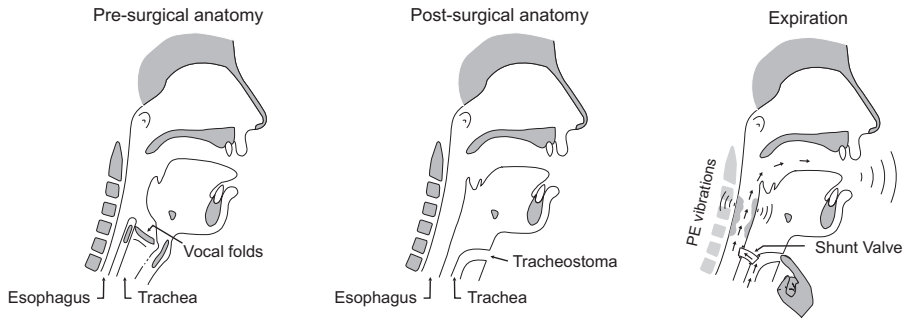

Fig. 1. Anatomy of a person with intact larynx *(left)*, anatomy after total laryngectomy *(middle)*, and the substitute voice *(right)* caused by vibration of the pharyngo-esophageal segment (pictures from [6])

quality of substitute voices is "low". Intercycle frequency perturbations result in a hoarse voice [1]. Furthermore, the change of pitch and volume is limited which causes monotone voice. Acoustic studies of TE voices can be found for instance in [2, 3]. In this paper, we will not concentrate on acoustic properties. The reduced sound quality and problems such as the reduced ability of intonation or voiced-voiceless distinction [4, 5] lead to worse intelligibility. For the patients this means a deterioration of quality of life, as they cannot communicate properly.
Another source of distortion is the so-called tracheostoma which is the upper end of the trachea (cmp. Fig. 1). In order to force the air to take its way through the shunt into the esophagus and allow voicing, the patient usually closes the tracheostoma with a finger. If the patient is not able to do this properly, loud "whistling" noises from the eluding air may occur.

In our work we examine how well TE speech is processed by a speech recognition system, how the recognizer can be adapted to TE voices and if the results can be used for evaluating the quality of a substitute voice automatically, i.e. if they correlate with experts' ratings. Initial results on these topics will be presented in the following.

2 The Baseline System

The speech recognition system used for the experiments was developed at our institute. It can handle spontaneous speech with mid-sized vocabularies up to 10000 words. The latest version is described in detail in [7].
For each frame a 24-dimensional feature vector which contains short-time energy, 11 Mel-frequency cepstral coefficients (MFCC) and their first-order derivatives are computed. The derivatives are approximated by the slope of a linear regression line over 5 consecutive frames (50 ms). The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The filterbank for the Mel-Spectrum consists of 25 triangle filters.

The system uses semi-continuous Hidden Markov Models (HMM). It models phones in a context as large as still statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones. The HMMs for each polyphone have three to four states. In the current experiments the codebook had 500 classes and a unigram language model was used, so that the results are mainly dependent on the acoustic models.

3 Training and Test Data

The baseline system for the experiments in this paper was trained with dialogues from the VERBMOBIL project [8]. The topic in the recordings is appointment scheduling. The data were recorded with a close-talk microphone at a sampling frequency of 16 kHz and quantized with 16 bit. The speakers were from all over Germany and thus covered most dialectal regions. They were, however, asked to speak standard German. About 80% of the 578 training speakers (304 male, 274 female) were between 20 and 29 years old, less than 10% were over 40. This is important in view of the test data, because the fact that the average age of our test speakers is more than 60 years may influence the recognition results.
11714 utterances (257,810 words) of the VERBMOBIL-German data (12030 utterances, 263,633 words, 27.7 hours of speech) were used for the training and 48 (1042 words) for the validation set. Thus we kept the same corpus partitions as in [7].

The test files were recorded from 18 male laryngectomees (64.2 ± 8.3 years old) with tracheoesophageal substitute speech. They had undergone total laryngectomy because of laryngeal or hypopharyngeal cancer at least one year prior to the investigation and were provided with a Provox® shunt valve. Each person read the story of "North Wind and Sun", a phonetically balanced text with 108 words (71 disjunctive) often used in German speaking countries in speech therapy. The duration of all 18 audio files together was 21 minutes, the test persons spoke 1980 words. In addition to the words of the text 32 different additional words were produced as reading errors. The vocabulary of the recognizer for the experiments consisted of the words occuring in the test data (71+32).
In order to get an age-matching set of normal laryngeal speakers, currently also a group of healthy older persons is being recorded.

4 Unsupervised Adaptation to Substitute Voices

The HMM interpolation technique was originally used for the sparse data problem. When a speech recognizer for a domain with a small amount of training data has to be built its acoustic models can be made more robust by interpolation with models from another recognizer. In [7] an interpolation method is described which was originally used to adapt a speech recognizer to non-native speech. In the experiments there each HMM has only one interpolation partner. In [9] an algorithm to select a variable number of partners was introduced. We combined the approach of [7] with the method described in [9] to adapt the

speech recognizer to substitute voices, but without using a second recognizer. First we converted the VERBMOBIL polyphone recognizer into a monophone recognizer. Nevertheless it still contained the original polyphones. These were now the candidates for the adaptation of the monophone models to TE speech. This was done unsupervised as follows:
With the original recognizer the best word chain was computed. It was assumed to be correct. Then the monophones underlying the best word chain were interpolated. First each monophone was interpolated with each single polyphone alone, i.e. the coefficients of the Gaussians for the two elementary HMMs were added with weighting factors that sum up to one. Remember that we use semi-continuous HMMs. For each monophone a set of n well fitting polyphones was chosen as interpolation partners then. The number n can be optimized in a separate step which will not be described here. For our experiments with the tracheoesophageal voices first one single interpolation partner was chosen for each HMM. Then, in a second step, the number of partners was set to 40, because this was the number that had achieved the best results in [9]. The interpolation weights were estimated using the EM algorithm [10].

The recordings from the 18 test speakers showed a wide range in intelligibility, volume, hoarseness of the substitute voices and sometimes also noises from the tracheostoma. Furthermore the data set was too small to be representative for all TE speakers and thus not suitable to be handled as a whole in a speech recognition task. Therefore interpolation was not done for the entire group of speakers in general, but to each single speaker separately which in principle lead us to 18 different recognizers. These in the following will be treated as one.
The differences in voice quality can clearly be seen in Fig. 2 where the recognition results are summarized. The worst speaker's word accuracy on the baseline polyphone recognizer ("baseline_poly") was only 2.7% while the best one reached 62.7%. The average value was 28.2% (see also Table 1) – a control group of 16 normal laryngeal speakers had shown an average of 83.7% (within a range from 75.0% to 93.5%). Then a monophone recognizer ("baseline_mono") was trained with the same VERBMOBIL data as the baseline system. We hoped that the more robust training of the monophones would have a positive effect on the recognition of the substitute voices. As the picture shows the "low quality" voices were recognized better while the monophone models were disadvantageous for the clearer voices. Thus the mean value rose only slightly to 28.7%. One outlier (speaker #10) appeared. The voice of this man had a gargling sound and he very often breathed hearable. It is not clear whether the reason for his bad results are connected to these facts. The interpolation of the monophone recognizer's HMMs with one ("interpol1") and 40 interpolation partners ("interpol40") enhanced the recognition for almost all speakers, where the latter approach with its mean word accuracy of 36.4% outperformed the former one by 3 percent points. Of course these results cannot be set in direct correlation to the baseline systems, because the new recognizers were optimized separately for each single speaker, but the results show that a high number of HMM interpolation partners seems to be better than a small one which is conform with [9].

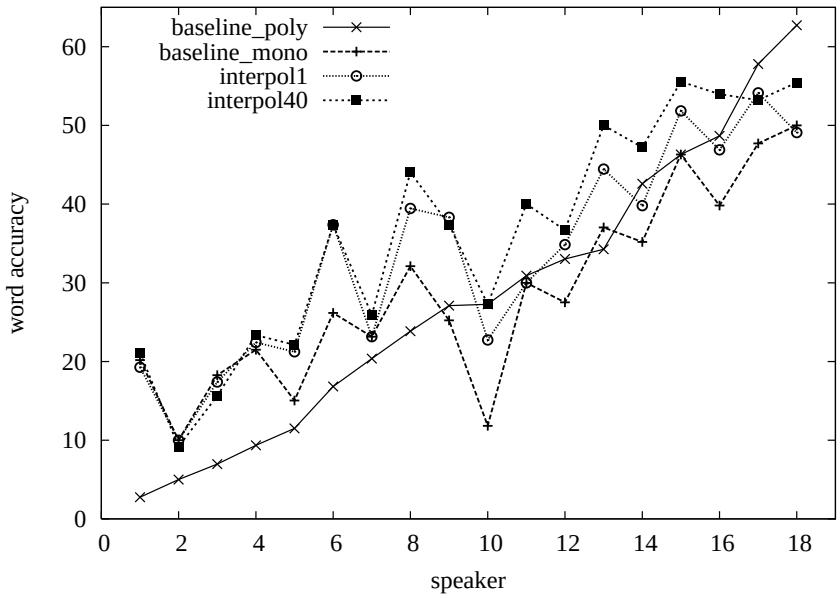

Fig. 2. Recognition results of four different automatic speech recognizers for 18 files with tracheoesophageal speech

Table 1. Average word accuracy for the used ASR systems; the interpolated recognizers were, however, optimized to a single person in an unsupervised manner and then evaluated on this particular person only

recognizer	baseline_poly	baseline_mono	interpol1	interpol40
mean WA	28.2	28.7	33.5	36.4
st. dev.	18.1	12.1	13.2	14.7

But this is not the only conclusion that can be drawn. The main outcome of the experiments is that speech recognition on tracheoesophageal voices is in principle possible, although for the lower quality voices more work has to be done.

In the next section we will discuss a basic approach for the comparison between the evaluation of a substitute voice by human raters and by an automatic speech recognition system.

5 Human and Automatic Intelligibility Rating

In speech therapy and rehabilitation a patient's voice has to be evaluated by the therapist. An automatically computed, objective measure would be a very helpful support for this task. In this section we present some initial experiments. At the Department of Phoniatrics and Paediatric Audiology at our university

Table 2. Correlation coefficients between single raters and the average of the 4 other raters for the criterion "intelligibility"

raters	all vs. K	all vs. L	all vs. R	all vs. S	all vs. U
correlation	+0.83	+0.82	+0.77	+0.85	+0.68

five experienced phoniatricians and scientific engineers evaluated the voices of the 18 test persons on criteria such as "hoarseness", "prosody" and "effort". Another criterion was "intelligibility". The scores given by the experts were represented by numbers between 1 (very high) and 5 (very low). It seemed to be obvious to us that a voice which is well intelligible for a human being will also achieve better results in automatic speech recognition (cmp. Section 4). So we chose this single criterion and compared the experts' rating to the word accuracy we got from our speech recognizer.

First of all we tested how homogeneous the expert group rated the test data. For the 18 files the correlation of each single rater's "intelligibility" scores to the average scores across the other four persons was calculated (compare Table 2). The two lowest correlation values were 0.68 and 0.77, the others were between 0.82 to 0.85. The inter-rater variance for the experts was 0.11. Then we measured the correlation between man and machine for the 18 recordings where the word accuracy across a speaker's entire utterance served as the automatically computed score. The results for the correlation to the average of the five experts are shown in Table 3. Considering the average of the raters, the best recognition systems for the task is the monophone recognizer with a correlation of -0.84. The coefficient is negative because high recognition rates came from "good" voices with a low score number and vice versa. The average score of the five raters and the word accuracy from the monophone recognizer are also depicted in Fig. 3. The baseline polyphone recognizer and the recognizer using 40 interpolation partners for each HMM reached a correlation of -0.83. The approach using the interpolation with only one partner was slightly worse (-0.81).
In a communication situation between humans the dialogue partners are able to adapt their hearing to the other person's voice. The same thing has been simulated by our HMM adaptation where the recognition system was always adapted to the particular person. Therefore these approaches will not be used in an objective evaluation method. Furthermore a polyphone recognizer is based on phonemes that have been spoken in a special context. If the evaluation of intelligibility is to be extended to free speech, there might be an influence on the result by the percentage of polyphones in the spoken text which are not included in the recognizer's inventory. For this reason the use of a monophone recognizer seems to be more advisable.

It is clearly visible that there's a strong correlation between the results of the human and the automatic analyzing method. This leads us to the assumption that the word accuracy will be very helpful as a part of a future automatic intelligibility or, in general, voice quality analyzer.

Table 3. Correlation coefficients between five human raters and the used ASR systems for the criterion "intelligibility"

rater	baseline_poly	baseline_mono	interpol1	interpol40
all	-0.83	-0.84	-0.81	-0.83

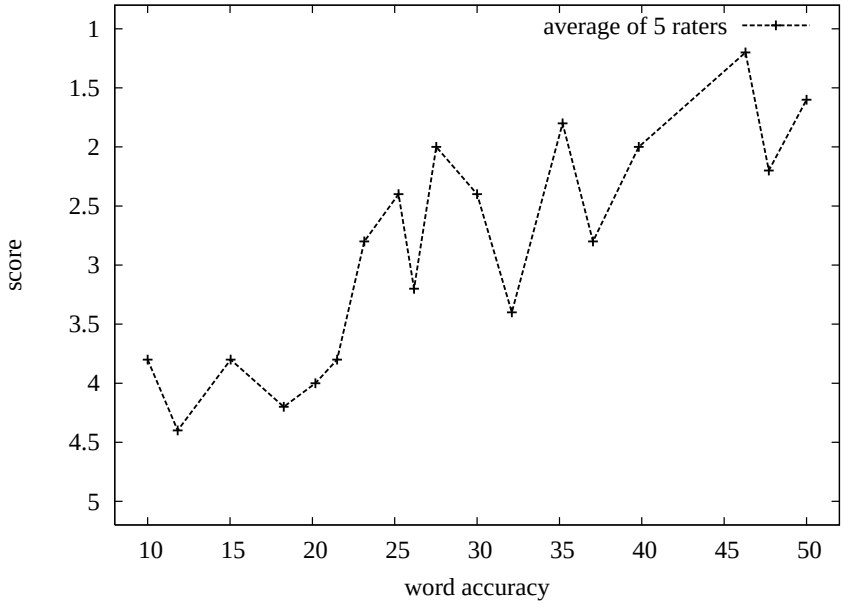

Fig. 3. Word accuracies vs. experts' scores for 18 patients with TE voice, scores of five experienced raters were averaged; the ASR system was a monophone recognizer

6 Conclusions and Outlook

A tracheoesophageal (TE) voice is a so-called substitute voice which is one possibility to give a patient back his ability to communicate by speech after laryngectomy. However, this voice which is produced in the pharyngoesophageal segment often shows low quality and intelligibility. We used unsupervised HMM interpolation to adapt a speech recognizer which was trained on non-pathologic voices to single recordings with TE speech. For 18 substitute voices an average word accuracy of 36.4% could be reached with 40 interpolation partners for each HMM. The baseline value had been 28.2%. The high error rates mainly arise from the fact that the speech recognizers were trained with normal, laryngeal speech. The training samples were mostly recordings from young people speaking standard German while the average age of the TE speakers was more than 60 years and some of them spoke dialect. More investigations have to be done with a bigger group of TE speakers which can also be interpolated as a whole.

In the field of voice evaluation we compared the intelligibility scores for recordings of TE voices from five experienced raters with the word accuracy from our system. The monophone recognizer's correlation was -0.84 on a standard text and thus showed that an automatic evaluation of the voice quality might be possible. In our current experiments the text reference for the calculation of the word accuracy was not the original text but a hand-labeled transcription of the audio files in order to exclude an influence of reading errors on the intelligibility evaluation. This ensured that the word accuracy reflects merely the acoustical recognition errors which was important for these basic experiments. Nevertheless the correlation between the word accuracies computed on the text reference and the experts' scores was also -0.84 for our data set. For a future clinical application the two sources of error have to be strictly divided. By the application of confidence measures and language models sections with reading errors could be detected in the recording. Then the remaining parts of the file will be used for the computation of the voice quality only.

References

[1] Schutte H.K., Nieboer G.J.: Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Filia Phoniatr Logop, **54** (2002) 8–18

[2] Robbins J., Fisher H.B., Blom E.C., Singer M.I.: A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. Journal of Speech and Hearing Disorders, **49** (1984) 202–210

[3] Bellandese M.H., Lerman J.W., Gilbert H.R.: An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. Journal of Speech, Language, and Hearing Research, **44** (2001) 1315–1320

[4] Gandour J., Weinberg B.: Perception of Intonational Contrasts in Alaryngeal Speech. Journal of Speech and Hearing Research, **26** (1983) 142–148

[5] Searl J.P., Carpenter M.A.: Acoustic Cues to the Voicing Feature in Tracheoesophageal Speech. Journal of Speech, Language, and Hearing Research, **45** (2002) 282–294

[6] Lohscheller J.: Dynamics of the Laryngectomee Substitute Voice Production. PhD thesis, Shaker-Verlag, Aachen, Germany (2003)

[7] Stemmer G.: Modeling Variability in Speech Recognition. PhD thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany (2004)

[8] Wahlster W. (ed): Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin (2000)

[9] Steidl S., Stemmer G., Hacker C., Nöth E., Niemann H.: Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In Michaelis B., Krell G. (eds): Pattern Recognition, 25th DAGM Symposium, Vol. 2781 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg (2003) 600–607

[10] Jelinek F., Mercer R.: Interpolated estimation of markov source parameters from sparse data. In Gelesma E.S., Kanal L.N. (eds): Proc. Workshop on Pattern Recognition in Practice. North-Holland, Amsterdam (1980) 381–397