ASR FOR ELECTRO-LARYNGEAL SPEECH

Anna K. Fuchs, Juan A. Morales-Cordovilla, Martin Hagmüller

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

anna.fuchs@tugraz.at, moralescordovilla@tugraz.at, hagmueller@tugraz.at

ABSTRACT

The electro-larynx device (EL) offers the possibility to re-obtain speech when the larynx is removed after a total laryngectomy. Speech produced with an EL suffers from inadequate speech sound quality, therefore there is a strong need to enhance EL speech.

When disordered speech is applied to Automatic Speech Recognition (ASR) systems, the performance will significantly decrease. ASR systems are increasingly part of daily life and therefore, the word accuracy rate of disordered speech should be reasonably high in order to be able to make ASR technologies accessible for patients suffering from speech disorders. Moreover, ASR is a method to get an objective rating for the intelligibility of disordered speech.

In this paper we apply disordered speech, namely speech produced by an EL, on an ASR system which was designed for normal, healthy speech and evaluate its performance with different types of adaptation. Furthermore, we show that two approaches to reduce the directly radiated EL (DREL) noise from the device itself are able to increase the word accuracy rate compared to the unprocessed EL speech.

Index Terms— Automatic Speech Recognition (ASR), electro-larynx (EL), speech enhancement, MLLR adaptation

1. INTRODUCTION AND RELATED WORK

The motivation to apply ASR on disordered speech is twofold. On one hand, ASR systems could be used to control assistive technologies whereas on the other hand, ASR systems can also be used for evaluation purposes. Based on the word accuracy rate speech intelligibility can be quantified. In existing ASR systems under controlled conditions, the word recognition accuracy is very high (around 90%). These ASR frameworks often comprise of a large amount of (continuous) speech. For disordered speech like dysarthric voice, where the ability to articulate is drastically reduced, building ASR systems is quite a problem, especially because the amount of speech material is much smaller than for normal speech. For patients with speech problems, speech recordings are significantly more exhausting and difficult than for normal speakers. State-of-the-art systems train triphone models and for this reason large amounts of speech material are needed. So far, ASR for disordered speech has been addressed by few authors, but it is an increasingly active research area. Some work is done on speech recognition of dysarthric speech which can have a very profound influence on speech intelligibility and thus, on the recognition results.

In [1], a database of dysarthric speech is used. This database is still much smaller than typical speech databases. The study investigates the influence of fundamental training and adaptation techniques on the ASR system where an improvement can be investigated. Different data sets (dysarthric and normal speech) and adaptation (MAP) on different targets were investigated. The speech material is comprised of 15 speakers with 250 unique words per speaker and approximately 50 minutes of speech per speaker. 12 PLP features are used to train the acoustic model. The best average word accuracy rate, of around 54%, was obtained by MAP adaptation of dysarthric speech model where the test speaker was also present in the training. Although the difference between normal speech an dysarthric speech is large, MAP can deal with it. Results strongly depend on the speaker and on the severity of the dysarthric speech.

In [2], the authors focused on speech material from patients suffering from head and neck cancer. A standard text read by 41 German laryngectomized (using tracheo-esophageal substitution voice) and 49 German patients who had suffered from oral cancer was evaluated. The results are compared to a control group of 40 speakers without speech pathology. The word recognition rate was then compared to perceptual ratings by a panel of experts. As an outcome it could be shown that ASR is a good measure with low effort to objectify and quantify intelligibility of disordered speech. Several language models were investigated. The ASR system was non-adapted. The results for the control group (76 ± 7) were significantly higher compared to the laryngectomized group (48 ± 19). The agreement, calculated using Spearman's correlation coefficient, between word recognition rate and the mean scores of the perceptual ratings is very high in both patient groups with -0.83 and -0.9 respectively.

In his doctoral thesis, Nakamura investigated a speech aid system for electro-laryngeal speech using statistical voice conversion [3]. Within this thesis he also carried out a case study of speech recognition for electro-laryngeal speech. He employed phonetically tied-mixture acoustic models. Maximum likelihood linear regression (MLLR) was the employed adaptation technique to transform the speaker independent model into a speaker dependent one. Two sets of speech data are used: 1) EL speech of a laryngectomized patient (native Japanese, 50 utterances for adaptation, 30 for test) and 2) speech of other types of speaking-impaired people (10 speaker (cerebral palsy, hearing-impaired,...)). The used speech material compromises of words, digits and short utterances. MFCC features are employed. For the second group around 20 to 40 utterances are taken for adaptation and around 20 for test. For 1) the accuracy for enhanced EL speech was almost 80%. The word accuracy for 2) was around 20% depending on the kind of disordered and increased to around 60% after the MLLR adaptation.

In this paper, we want to use a parallel electro-laryngeal, healthy speech database for evaluation using a ASR system. We want to show that EL speech is more applicable to ASR than dysarthric speech because the nature of the distortions for these two disordered speeches are different. We also want to find out whether EL speech enhancement approaches can be evaluated using ASR in terms of

This work has been supported by HEIMOMED Heinze GmbH & Co.KG and partially funded by the DIRHA European project FP7-ICT-2011-7-288121.

intelligibility. Compared to [3] we want to use a larger and more balanced database and we want to focus on the adaptation possibilities.

2. EXPERIMENTAL SETUP

2.1. Database Description

The speech material originates from the German parallel ELHE database and consists of up to 500 different utterances. Each utterance was spoken one time with healthy speech (HE) and one time with the EL device (EL) in order to compare differences between healthy and disordered utterances. According to [4], who carried out listening tests, there are no significant perceptual differences between EL speech produced by a patient or by a healthy subject. The utterances have been recorded in sessions and within each session a well known text, "Der Nordwind und die Sonne", has been recorded in 6 separate utterances. The speech material consists of phonetically rich utterances from different German speech corpora. All in all the subjects had to read up to (two times) 503 utterances. In total this database consists of 5024 utterances. Descriptive statistics about the parallel ELHE database can be seen in table 1. The utterances per speaker contain 2983 words. Without counting multiple occurrences there are 1439 words. 1091 words only occur once.

ID	Age		# Sentences	Length	μ_{f_0}	σ_{f_0}
F01	28	EL	503	45min28s	192	7
		HE	503	29min57s	198	27
F03	31	EL	250	19min51s	199	6
		HE	250	13min48s	175	28
M02	38	EL	503	36min30s	99	4
		HE	503	24min55s	113	17
M04	50	EL	503	52min10s	93	1
		HE	503	30min5s	140	30
M05	29	EL	503	45min56s	93	0
		HE	503	26min02s	138	28
M06	29	EL	250	19min32s	94	1
		HE	250	12min58s	119	20
Sum			5024	5h57min12s		

Table 1. *Number of utterances in the parallel ELHE database; Mean value of f_0 - μ_{f_0} and standard deviation σ_{f_0}.*

The Austrian German native speakers have been healthy subjects with an average age of 29.5 years (female) and 36.5 years (male). The subjects used a Servox Digital. Two female (F01 and F03) and four male speaker (M02 and M04, M05, M06) have been recorded. More male speakers than female speakers have been recorded because this represents the statistics of EL patients. The fundamental frequency of the device was adjusted to a comfortable level for each speaker separately. 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. All recordings were carried out on-site at the recording studio of the Signal Processing and Speech Communication Laboratory at Graz University of Technology. 445 (192) utterances per speaker compose a phonetically balanced set for training and 58 utterances per speaker for testing. The number of words of the training is 579 and for the test 2404.

Additionally, around 2500 clean utterances of the Bavarian Archive for Speech Signals (BAS) PHONDAT-1 [5] database sampled at 16 kHz were used. These utterances correspond to 25 different speakers from both genders resulting in around 100 utterances per speaker. These subjects were native German speakers and used the same speech material as the speakers of the parallel ELHE speech database.

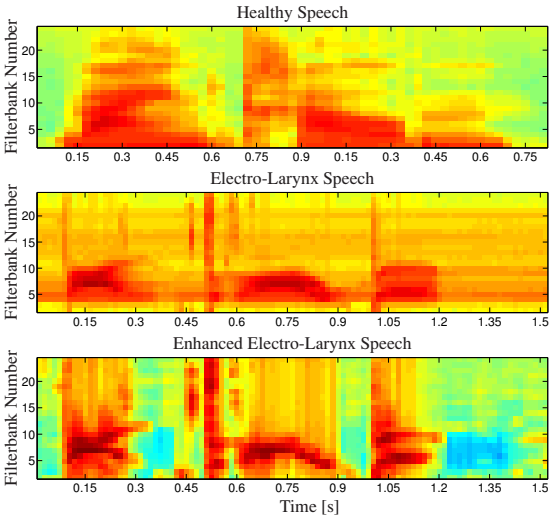

Fig. 1. *The log mel-filterbank spectra of the word "Weintrauben"; Healthy (HE) speech (upper plot), Electro-Larynx (EL) speech (middle plot) and enhanced EL speech using modulation filtering (MF) (lower plot).*

2.2. EL Speech Enhancement Strategy

Listening tests carried out by [6] have shown that EL speech can be most improved by removing the directly radiated EL (DREL) sound and providing pitch information. In this paper, we use two simple enhancement strategies to reduce the DREL sound: 1) spectral subtraction (SS) and 2) modulation filtering (MF).

The first method is SS. The DREL noise of an EL is only slowly varying and therefore [7] applied SS to EL speech. SS is based on estimating the noise power spectrum and then subtracting this spectrum from the signal power spectrum. Although SS suffers from the problem that the direct noise is synchronized with the tract excitation, and additionally, that environmental background noise and the directly radiated EL noise have completely different properties, this method was able to reduce the DREL to a large extent.

The second method, MF, filters out the DREL sound in the modulation frequency domain. This approach introduced by [4] takes advantage of the different properties of the EL speech sound and the DREL sound. As the directly radiated component of the EL energy is not modulated by the articulatory organs, but transmitted over the air to the human ear on a direct path, this signal is only modulated at a very low frequency and can effectively be assumed to be time-invariant. If we consider that the speech sound is a time and frequency dependent modulation of the excitations signal – in our case the EL sound – then we only have to suppress the signal path which is constant. To do so, a notch filter is placed at a modulation frequency of $f_n = 0$ Hz.

The log mel-filterbank spectra of HE speech, EL speech and enhanced EL speech using MF is illustrated in Fig. 1. It can be seen that there is a mismatch between the HE and EL domain in terms of high- and low-frequency deficit as well as differences in the position, the bandwidth and the energy of the formants [8].

Speaker ID	Baseline [%]						Speaker MLLR Adaptation[%]			Domain MLLR Adaptation[%]
	H_H	H_E	E_{E1}	E_{E2}	EH_{E1}	EH_{E2}	H_E	E_{E2}	EH_{E2}	dHE_{E2}
F01	97.39	5.22	15.36	78.55	27.25	67.25	51.50	81.74	83.77	81.16
M02	98.89	28.61	64.71	91.45	83.33	88.70	61.40	91.32	89.72	88.44
F03	99.56	0.22	31.43	62.58	39.00	55.31	1.64	80.71	71.68	54.68
M04	99.27	-6.57	47.93	84.67	37.24	71.83	63.04	84.43	80.66	85.40
M05	98.95	3.66	42.93	75.39	57.74	70.03	37.40	81.15	76.70	80.63
M06	99.33	5.37	55.96	83.20	45.03	75.43	15.82	89.19	87.70	83.67
Average	98.96	5.53	42.62	79.31	48.02	70.84	36.82	84.76	81.70	79.00

Table 2. *Results of ASR for different setups; 1) Baseline, 2) Speaker adaptation using MLLR and 3) Domain adaptation using MLLR; H_H and H_E – Training: healthy, Test: healthy, electro-laryngeal; E_{E1} and E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal (1 - speaker is not included in training, 2 - speaker is included in training); EH_{E1} and EH_{E2} – Training: mixed healthy and electro-laryngeal, Test: electro-laryngeal (1 and 2 as before); dHE_{E2} – Training: healthy, Adaptation: to electro-laryngeal domain, Test: electro-laryngeal.*

Speaker ID	Domain MLLR Adaptation[%]					
	dHE_{E1} – Adapted material					
	F01	M02	F03	M04	M05	M06
F01	78.59	-36.86	0.89	-17.07	-0.91	0.88
M02	10.03	82.72	0.28	8.17	3.10	0.28
F03	4.41	-1.81	16.56	2.61	10.68	0.00
M04	-18.97	-33.33	1.01	82.43	-3.89	-6.72
M05	3.14	-29.92	5.65	-3.48	73.82	0.52
M06	-2.75	0.93	0.00	8.47	2.26	27.29

Table 3. *Results of ASR for the setup: Domain adaptation using MLLR; dHE_{E1} – Training: healthy, Adaptation: to speaker dependent electro-laryngeal domain, Test: electro-laryngeal.*

2.3. Automatic Speech Recognition System

Although it is necessary to estimate a large number of parameters compared to monophone HMMs, in this paper we built an ASR system based on HMM triphones. The advantage is that the size of the lexicon can easily be increased in the future and thus, is most useful in real applications. Also, the characteristics of human voices is reasonably well expressed with triphones. Both the Front-End (FE) and the Back-End (BE) have been derived from the standard base-line recognizer employed in Aurora-4 database [9]. The most important parameters of the FE are: 32 ms frame length and 100 Hz frame rate; 26 triangular filters for the Mel-spectrum; 13 Mel-Frequency Cepstral Coefficients (MFCCs) and cepstral mean normalization (CMN). Delta and delta-delta features with a window length of 5 (half length 2) are also appended, obtaining a final feature vector with 39 components.

To train the triphones, the BE employs a transcription of the training corpus based on 34 SAMPA-monophones. This transcription has been derived from a more detailed monophone transcription (based on 44 SAMPA-monophones) by means of a careful clustering of the less common monophones. Each triphone is modeled by a hidden Markov model (HMM) of 6 states and 8 Gaussian-mixtures/states. By means of a monophone classification created with the help of a linguistic, a tree-based clustering of the states is also applied to reduce the complexity and the lack of training data. Tree-based clustering also allows the creation of triphone models which have not been observed in the training stage. In this paper we use a bigram language model. The perplexity of the language model is around 3.5. The higher the value of the perplexity is, the worse the prediction in the test set is and the worse the recognition results are. In [10] the authors predict a perplexity of around 131 for bigram language model using the 'Sherlock Holmes' books.

One of the most powerful and popular adaptation techniques is maximum likelihood linear regression (MLLR) [10]. In this paper, we apply two kinds of adaptation: **1) Speaker dependent MLLR adaptation** based on a class tree regression and **2) Domain MLLR adaptation** based on retraining the model using new data from the target domain.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Results are presented using the word accuracy rate. According to [10] the percentage accuracy is defined as

$$W_{Acc} = \frac{N - S - D - I}{N} \cdot 100\%$$

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

Three types of sets are used to train the triphones: 1) speech data from healthy individuals (H_H or H_E), 2) electro-laryngeal speech data (E_{E1} or E_{E2}) and 3) a mixture of healthy and electro-laryngeal speech data (EH_{E1} or EH_{E2}).

Using this notation, H stands for healthy, and E for electro-laryngeal. The capital letter indicates the training, and the subscript indicates which data is tested. The difference in the subscript between E_{E1} and E_{E2} (HE_{E1} and HE_{E1}) indicates whether the speech material of the tested speaker occurs in the training (1 if it does not, 2 if it does). It must be noted that the amount of training material differs for the different types and speakers. For training and test of speakers F01, M02, M04 and M05, around 500 utterances are available and for F03 and M06 only around 250.

Speaker ID	Baseline [%]			Domain MLLR adaptation [%]		
	E_{E2}	$E_{E2_{SS}}$	$E_{E2_{MF}}$	dHE_{E2}	$dHE_{E2_{SS}}$	$dHE_{E2_{MF}}$
F01	78.55	89.86	87.25	81.16	88.99	83.48
M02	91.45	93.61	94.72	88.44	88.89	92.50
F03	62.58	62.75	60.78	54.68	53.16	51.85
M04	84.67	85.16	89.54	85.40	83.94	88.70
M05	75.39	83.25	84.03	80.63	78.01	82.46
M06	83.20	82.77	92.17	83.67	75.17	89.49
Average	79.31	82.90	84.75	79.00	78.03	81.41

Table 4. *Results of ASR for different setups; 1) Baseline and 2) Domain adaptation using MLLR ; E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal (2 - speaker is included in training); $E_{E2_{SS}}$ – enhanced electro-laryngeal speech using spectral subtraction (SS); $E_{E2_{MF}}$ – enhanced electro-laryngeal speech using modulation filtering (MF); dHE_{E2} – Training: healthy, Adaptation: to electro-laryngeal domain, Test: electro-laryngeal; $dHE_{E2_{SS}}$ – enhanced electro-laryngeal speech using SS; $dHE_{E2_{MF}}$ – enhanced electro-laryngeal speech using MF.*

3.1. Experiments on EL speech

In table 2, word accuracy (W_{Acc}) rates are shown for the different setups. The W_{Acc}, when training material only consists of healthy speech (BAS as well as healthy speech material from the 6 speakers) and we test on healthy speech, is 98.96% (Baseline - H_H). When the test is carried out on electro-laryngeal speech, the performance is very low (5.53%; Baseline - H_E) due to the mismatched domain (differences in the position, the bandwidth and the energy of the formants between healthy and electro-laryngeal speech – see also Fig. 1).

The performance of speaker M02 is consistently good for each setup. Even in the mismatched domain (training: healthy; test: electro-laryngeal) this speaker performs well (28.61%; Baseline - H_E). This speaker is most used to handle the EL device. Speaker F03 performs worse than anybody else. Informal listening tests verified that this speaker is less intelligible than the others when speaking with the EL device. This is one reason of her low performance. Another reason is that this speaker is female, and female speakers are less represented than male speakers in the parallel ELHE database. The same case holds for speaker F01 in the baseline experiments for E_{E1} (15.36%).

When speech material of electro-laryngeal speech is added to the healthy training, (EH_{E1} and EH_{E2}) the word accuracy rate improves to 70.84% regarding H_E. The results improve even further using only electro-laryngeal speech for training (79.31%; Baseline - E_{E2}). Considering that only 2164 utterances are used in the training this result is good due to the low perplexity of the grammar (see subsection 2.3).

Using speaker MLLR adaptation, results for the healthy-disordered mixed training (EH_{E2}) reach a value of 81.70%, and increase to 84.76% for the training with only electro-laryngeal speech (E_{E2}). This shows that electro-laryngeal speech is sensitive to the speaker change.

In the next experiment we investigate the case when only little data is available. Connected to that we also want to show that it is possible to obtain a robust electro-larygneal model starting from a healthy speech model. For this reasons we apply domain MLLR adaptation of healthy speech to electro-laryngeal speech. With this approach we reach a word accuracy of 79.00%, which is in the same order as the electro-laryngeal speech model (E_{E2}). Additionally, we applied domain MLLR adaptation to a specific speaker. Looking at the results in table 3 we can see that only the speakers included in training also perform well in the test. Also the baseline results of E_{E1} (42.62%) and E_{E2} (79.31%) of table 2 confirm this circumstance. These results show that the EL models for ASR are strongly speaker dependent probably due to the different ways to articulate EL speech.

3.2. Experiments on enhanced EL speech

For these experiments, we applied two basic enhancement strategies, explained in subsection 2.2. These strategies are tested on the E_{E2} model. Results can be seen in table 4. Electro-laryngeal speech is first enhanced using the two strategies, then the models are trained and tested using these signals (Baseline - E_{E2}, $E_{E2_{SS}}$, $E_{E2_{MF}}$). For the domain MLLR adaptation, we take the enhanced speech utterances to adapt to the healthy speech model (dHE_{E2}, $dHE_{E2_{SS}}$, $dHE_{E2_{MF}}$). We can observe that both enhancement algorithms improve the results regarding the baseline E_{E2}. In general MF outperforms SS because the multipath approach to reduce the DREL noise reflects the true nature of DREL noise better.

For the domain MLLR adaptation results, the changes in the average word accuracy rate are -0.7% and 2.41% regarding $dHE_{E_{E2}}$. This suggests that the domain adaptation can deal with EL speech as well as enhanced EL speech.

4. CONCLUSION

In this paper, we investigated the behavior of ASR systems with disordered speech, namely electro-laryngeal, for training and testing which introduced a new kind of disordered speech to ASR. The speech material originates from the parallel ELHE database which has been recorded in our laboratory and consists of four male and two female speakers.

One important conclusion is that the recognition results for electro-laryngeal speech are admissible because people tend to articulate very clearly in order to be understandable and because the stationary noise of the electro-larynx device can be modeled by the ASR training. Another important conclusion is that the EL model for ASR strongly depends on the speaker and with a speaker dependent MLLR adaptation strategy, electro-laryngeal speech results are nearly as high as for healthy speech. Although there is a large mismatch between the two domains, as soon as we include electro-laryngeal speech in the training, the ASR system performs well. Also we have seen that the performance of the female speaker is lower due to the dominance of the male speakers in the database. Furthermore, recognition results could be improved by using simple

speech enhancement strategies which suggests that ASR can be used to evaluate the intelligibility of enhanced electro-laryngeal speech.

We have successfully applied electro-laryngeal speech to an ASR system and achieved high word accuracy rates. All of this leads to the conclusion that if some preprocessing is done, EL users can have access to ASR technologies.

5. REFERENCES

[1] Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012.

[2] Andreas Maier, Tino Haderlein, Florian Stelzle, Elmar Nöth, Emeka Nkenke, Frank Rosanowski, Anne Schützenberger, and Maria Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 1:1–1:7, Jan. 2010.

[3] Keigo Nakamura, *Speaking-aid Systems Using Statistical Voice. Conversion for Electrolaryngeal Speech*, Ph.D. thesis, Nara Institute of Science and Technology, 2010.

[4] Martin Hagmüller, *"Speech Enhancement for Disordered and Substitution Voices"*, Ph.D. thesis, Graz University of Technology, 2009.

[5] F. Schiel and A. Baumann, "Phondat 1, corpus version 3.4," 2006.

[6] Geoffrey S Meltzner and Robert E Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech.," *J Speech Lang Hear Res*, vol. 48, no. 4, pp. 766–79, 2005.

[7] David Cole, Sridha Sridharan, Miles Moody, and Shlomo Geva, "Application of noise reduction techniques for alaryngeal speech enhancement," in *Proc. IEEE TENCON'97*, Brisbane, Australia, Dec. 1997, pp. 491–494.

[8] Yoko Saikachi, *Development, perceptual evaluation, and acoustic analysis of amplitude-based F0 control in Electrolarynx speech*, Ph.D. thesis, Harvard-MIT Division of Health Sciences and Technology, 2009.

[9] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0," Etsi stq-aurora dsr working group, November 19 2002.

[10] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.