interspeech 2008 incorporating SST 08

Longitudinal study of ASR performance on ageing Voices

Ravichander Vipperla, Steve Renals, Joe Frankel

The Centre for Speech Technology Research, School of Informatics
Edinburgh University, UK

Abstract

This paper presents the results of a longitudinal study of ASR performance on ageing voices. Experiments were conducted on the audio recordings of the proceedings of the Supreme Court Of The United States (SCOTUS). Results show that the Automatic Speech Recognition (ASR) Word Error Rates (WERs) for elderly voices are significantly higher than those of adult voices. The word error rate increases gradually as the age of the elderly speakers increase. Use of maximum likelihood linear regression (MLLR) based speaker adaptation on ageing voices improves the WER though the performance is still considerably lower compared to adult voices. Speaker adaptation however reduces the increase in WER with age during old age.

Index Terms: Ageing Voices, longitudinal study, SCOTUS corpus, MLLR

1. Introduction

With ageing, human bodies undergo several degenerative changes. Changes also occur in the respiratory system, larynx and the oral cavity which form the human speech production mechanism [1]. Significant changes in these organs affecting speech include loss of elasticity in the respiratory system leading to decreased lung pressure, calcification of the laryngeal tissues leading to the instability of the vocal fold vibrations, loss of tongue strength, tooth loss and changes in dimensions of the oral cavity. Studies reported in [2] indicate that there is a difference in several acoustic parameters between aged voices and adult voices. Aged voices are typically characterised by changes in fundamental frequencies, increased breathiness, increased jitter and shimmer, and slower speaking rate. These changes seem to affect the ASR recognition accuracies as well.

Results published in [3] and [4] show that the automatic speech recognition (ASR) accuracies for elderly voices above 70 years of age are significantly lower than those of adult speakers. Acoustic models trained purely on elderly speech were used in [4] and [5], however the improvement was limited.

It is not clear if the recognition accuracies drop due to certain medical conditions that the speakers may have during old age, or if the drop in performance is just a natural affect with ageing. It is therefore of interest to perform an independent set of experiments on healthy individuals aver 60 years of age. It is also of interest to see how the ASR accuracy varies as people grow older after 60 years of age, especially to see if there is a sudden drop in performance at certain age or if the change is gradual. In this paper, we have used the voices of 7 United States Supreme court judges from the SCOTUS corpus who are in good health condition and whose speech data is available over several years to perform a longitudinal study.

We have also performed experiments to understand if MLLR based speaker adaptation [6] can be used to eliminate or reduce the age related effects on ASR performance.

In the following section, we briefly describe the SCOTUS corpus. In section 3, the experimental setup is described and the results of our experiments are presented, followed by a brief discussion in section 4.

2. SCOTUS Corpus

The SCOTUS speech corpus is the collection of the audio recordings of the proceedings of the Supreme court of the United States. These recordings have been made public under the Oyez project at http://www.oyez.org. Each recording's duration is about one hour and consists of speech from the advocates and judges arguing the case. These recordings were archived on reel-to-reel tapes, which were later digitised and made public.

Though the recordings from the 1980s to the present date are currently available online, complete transcripts with speaker information are available from the later half of 1990s. Hence only those files with speaker tags were used in our experiments. In all, the experimental corpus contains 534 recordings. It consists of speech from 10 Judges over several years and about 500 advocates. The birthdates of the Judges are known and hence their age at the time of an argument can be precisely calculated. The birthdates of the advocates are not easily available, hence their ages are being approximated by using the year of their law graduation and assuming their age at graduation to be 25. For the purpose of the current experiments, we have made a fair assumption that they are all between 25 and 55 years of age and have used their speech as adult voices.

In order to obtain the sentence boundaries and speaker turn alignments in each recording, forced alignment was performed on each recording using acoustic models trained on continuous telephone speech (CTS) and adapted on meetings data recorded by the International Computer Science Institute (ICSI) and at the National Institute of Standards and Technology (NIST).

To the best of our knowledge, SCOTUS corpus has only been used in speaker identification experiments [7] previously and there have been no published ASR baseline results.

3. Experiments

The goals of the experiments are: 1) To verify if the ASR performance on the ageing voices is inferior to that of adult voices 2) To understand how the ASR performance varies with increasing age during the old age. and 3) To understand if speaker adaptation using MLLR reduces any age related effects on ASR performance.

3.1. Experimental Setup

The SCOTUS corpus in MP3 format was first converted to 16KHz wav format and then parametrised using perceptual linear prediction (PLP) Cepstral features. A window size of 25ms

Accepted after peer review of full paper

and frame shift of 10ms were used for feature extraction. Energy along with 1^{st} and 2^{nd} order derivatives were appended giving a 39-dimensional feature vector.

To train the acoustic models, 90 hours of speech data from advocates was used. A significant portion of the entire corpus is from males, hence the training data set is also similarly skewed in favour of males with around 77 hours of speech from males and 13 hours of speech from females. The acoustic models have been trained as crossword context dependent triphone hidden markov models (HMM).

To construct the test vocabulary and the language model, transcripts of the SCOTUS corpus from the 1980s and the transcripts of the training data were used. The test vocabulary consists of 28818 words. A back-off bigram language model was constructed for test set decoding.

3.2. ASR performance on Adult vs Elderly speakers

In this experiment, we compare the overall WER on adult voices and elderly voices. The adult test set consists of 8655 utterances from 125 speakers, out of which 100 speakers are male and 25 female. There is no overlap of speakers in the adult test set and the training set. For the elderly voices test set, data from 7 out of the 10 judges available was used, as data was available for those speakers over a number of years, and they were all well above 60 years of age. 5 of these 7 speakers are males and 2 are females. The elderly test set includes around 200 utterances for each speaker from each year starting from 1999. The total number of utterances in this set is 11884 out of which 3183 utterances are from females and 8701 utterances are from males. Table 1 shows the comparison of the word error rates of adult test set and elderly test set.

Table 1: *Comparison of WER on Adult and elderly voices.*

Word Error Rate (WER) %		
	Adult Voices	Elderly voices
Overall	36.4	47.8
Male	34.6	43.2
Female	46.8	61.2

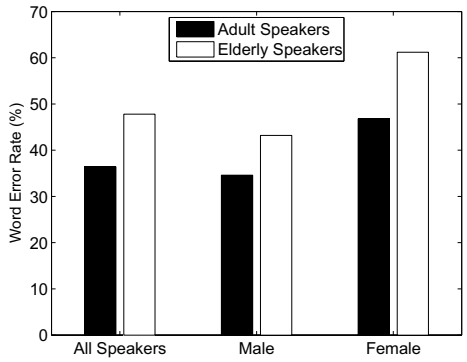

Figure 1: *Comparison of WER on Adult and elderly voices*

3.3. Longitudinal ASR performance on elderly voices with increasing age

In this experiment, the test set used is the elderly test set as described in the previous experiment. A set of 200 utterances from each year for each speaker were decoded and the ASR WER results are shown in Table 2. Speakers with IDs 02,03,04,05 and 08 are males while speakers 07 and 10 are females. The results have been plotted in Figure 2. A least squares line fit for the WER results of each speaker is also plotted to visualise the trend. It can be seen that though there are small fluctuations in WER over each year, there is a general pattern in the results showing increase in WER gradually as the age increases.

Table 2: *WER with increasing age on elderly voices.*

Word Error Rate (WER) %							
	Speaker ID						
Age	02	03	04	05	07	08	10
59				37.0			
60				40.5		37.4	
61				38.2		38.7	
62			45.3	36.8		39.1	
63		40.1	45.9	37.7		39.6	
64		41.4	48.3	38.0		40.0	
65		41.1	49.4	38.1		42.7	
66		41.1	48.7	37.3	61.5	40.1	
67		43.7	52.3	38.6	63.3	41.4	
68		41.9	50.8		67.4	41.7	
69		45.0	49.3		66.2		47.7
70		41.0	52.0		68.8		49.2
71		43.0			72.7		45.2
72					66.8		50.6
73					69.7		49.1
74					73.9		50.8
75							50.5
79	49.3						
80	48.9						
81	47.9						
82	52.1						
83	46.4						
84	48.9						
85	48.7						
86	50.1						
87	55.1						

3.4. Experiments using MLLR speaker adaptation

In this set of experiments, the test sets for adult and elderly voices remain the same as previous experiments. For computing the MLLR regression matrices for each adult speaker, a development set comprising about 20-250 utterances (average 70 utterances) available for each test speaker was used. For each elderly speaker, about 300 utterances were used to compute speaker regression matrices.

For each speaker, MLLR transform matrices were computed for mean adaptation. The regression class tree consisted of 2 classes, one for speech and one for non-speech.

Table 3 summarises the WER results on adult voices and elderly voices using MLLR. Use of speaker adaptation is seen to bridge the gap between adult and elderly speech from 11.4% to 7.7%

Table 4 summarises the results of the longitudinal study

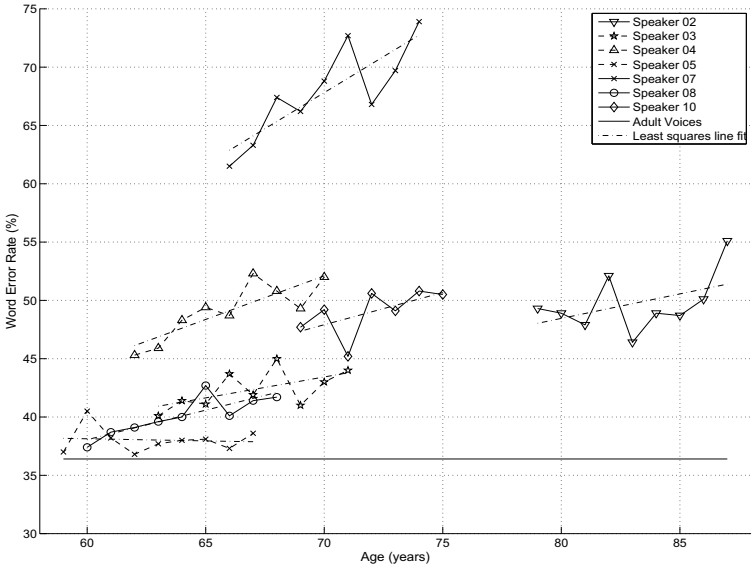

Figure 2: *WER with increasing age on elderly voices.*

Table 3: *Comparison of WER on Adult and elderly voices using MLLR.*

Word Error Rate (WER) %		
	Adult Voices	Elderly voices
Overall	33.1	41.8
Male	32.9	41.0
Female	33.8	44.0

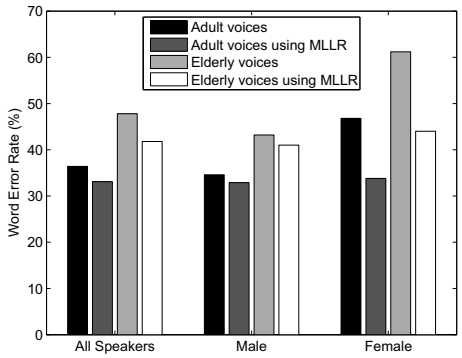

Figure 3: *Comparison of WER on Adult and elderly voices using MLLR*

with speaker adaptation for the 7 test elderly speakers. The longitudinal plots for each speaker are shown in Figure 4. A least square line fit for the WER over several years for each speaker are also plotted to understand the longitudinal performance.

Table 4: *WER with increasing age on elderly voices using MLLR*

Word Error Rate (WER) %							
	Speaker ID						
Age	02	03	04	05	07	08	10
59				35.6			
60				36.6		37.3	
61				35.4		37.3	
62			45.7	34.1		37.7	
63		38.9	44.0	35.0		37.7	
64		39.3	46.7	34.7		38.7	
65		38.7	48.0	35.7		41.0	
66		41.1	47.1	35.3	43.3	38.3	
67		40.1	48.8	36.2	45.8	39.7	
68		41.5	49.0		44.5	40.0	
69		38.6	46.7		49.9		38.2
70		39.5	50.2		45.8		39.8
71		39.8			47.9		37.8
72					44.4		39.5
73					46.7		41.7
74					49.0		40.2
75							41.0
79	46.9						
80	46.3						
81	45.3						
82	49.1						
83	44.4						
84	46.0						
85	47.0						
86	47.7						
87	51.0						

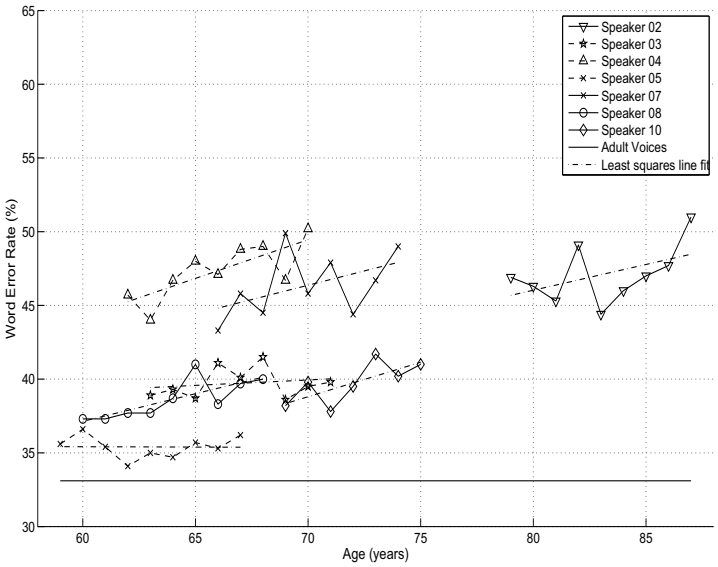

Figure 4: *WER with increasing age on elderly voices using MLLR transforms*

4. Discussion

From the results in Table 1, it is seen that the overall WER increases by 11.4% absolute or 31.3% relative from adult voices to elderly voices. For the male voices, WER increases by 8.6% which is consistent with the results published in [4] and [3]. For the female speakers, though the increase in WER is over 14.4% absolute, the figures may not be clearly indicative of the performance of the ageing female voices due to small sample test set and insufficient female speech in the training set.

From Figure 2, the longitudinal study results indicate that the WER gradually increases with age during old age. Since the number of utterances for each year for each speaker was limited, variations in the WER are expected, however the least squares line fit for WER of all the speakers have a positive slope which suggests an increase in WER with age especially after 65 years of age. For speakers 4, 7 and 8, F-tests show that there is strong evidence ($p < 0.01$) of a linear trend, and we conclude that in these cases WER is indeed increasing with age.

In the regression plots, a significant pattern with increasing age is not seen. This suggests that there is no clear correlation between chronological ageing and voice ageing across speakers.

Use of MLLR speaker adaptation decreases the difference between adult and elderly speech to 8.7% absolute or 26.3% relative. Adaptation narrows the gap between adult and elderly speech as seen from Figure 3. Longitudinal studies on elderly voices using MLLR adaptation also show a gradual increase in WER with age. For the case where MLLR has been used, we find that only speakers 4 and 8 show statistically significant evidence of a linear trend of increasing WER with age. In this case, the application of MLLR is shown to reduce the effect of increasing WER for speaker 7. The slopes of the longitudinal plots of each speaker using MLLR adaptation are less than those without adaptation, indicating that speaker adaptation can reduce the age related effects to some extent.

5. Conclusion

In this paper, we have investigated the performance of ASR on ageing voices. We found the WER rates for elderly voices to be significantly higher than adult voices. The WER for ageing voices increases gradually with age. Use of speaker adaptation by MLLR improves the performace for elderly voices but cannot achieve the same performance as adult speech. However we can reduce their deterioration with age through adaptation. Future work will be to develop adaptation methods which are particularly suited to reducing the performance gap between ASR on ageing and adult voices.

6. Acknowledgement

This study was supported by MATCH grant, funded by the Scottish Funding Council. The authors wish to thank Prof. Jerry Goldman, Prof. Mark Liberman and Dr. Jiahong Yuan for their valuable help in the SCOTUS corpus experimental setup.

7. References

[1] Linville, S. E., "Vocal aging", Singular Thomson Learning, 2001

[2] Xue, S, A., and Hao G. J., "Changes in human vocal tract due to aging and the acoustic correlates of speech production: a pilot study", Journal of Speech, Language and Hearing Research, 46(3):689-701, 2003

[3] Wilpon, J. G., and Nacobsen, C. N., "A study of speech recognition for children and the elderly", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, IEEE Press: 349-352, 1996.

[4] Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and Shikano, K., "Acoustic Models of the Elderly for Large-Vocabulary Continuous Speech recognition", Electronics and Communications in Japan, Part 2, 87(7): 49-57, 2004

[5] Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., and Levin, B., "Recognition of elderly speech and voice driven document retrieval", Proc IEEE International Conference on Acoustics Speech and Signal Processing, 1:145-148, 1999

[6] Gales, M. J. F., and Woodland, P. C., "Mean and Variance Adaptation Within the MLLR Framework", Computer Speech and Language, 10:249-264, 1996

[7] Yuan, J., and Liberman, M.,"Speaker identification on the SCOTUS corpus", Accepted for proceedings of ASA 2008