Automatic Speech Recognition and Error Analyses of Dutch Oral Cancer Speech

Kirsten Wildenburg[1], Bence M. Halpern[2,3,5], Teja Rebernik[1,3], Thomas Tienkamp[1], Rob J.J.H. van Son[2,3], Vass Verkhodanova[1], Max J.H. Witjes[1,4] and Martijn Wieling[1]

[1]University of Groningen, [2]University of Amsterdam, [3]Netherlands Cancer Institute, [4]University Medical Center Groningen, [5]Delft University of Technology

MOTIVATION

- ASR has significantly improved due to the introduction of deep learning [1]
- However, oral cancer speech is recognized poorly by ASR systems [2]
- It is important to develop ASR systems specifically for oral cancer speech to ameliorate patients' quality of life [3]
- A phoneme-level error analysis could potentially guide future ASR development

RQs AND HYPOTHESES

- RQ1 *'What **phonemes** in **oral cancer speech** cause **higher recognition error rates** in a standard **ASR** system compared to healthy speech in **Dutch**?'*
  → Plosives, alveolar sibilants and certain vowels are expected to cause higher error rates [2,4,5]
- RQ2 *'Does the **surgical treatment** of oral cancer patients **influence** the **ASR performance** on oral cancer speech?'*
  → Mandibulectomy is expected to impact ASR performance more than for glossectomy [4,6,7]

METHODS

1. NKI-UMCG-RUG oral cancer speech corpus: N = 11 (oral cancer: 6; control: 5)
   a. 3 mandibulectomy patients
   b. 3 (partial) glossectomy patients
2. Data preprocessed with librosa Python library [8]
3. Data run through ESPnet [9] with a Dutch pre-trained Conformer [10] model
4. Extensive error analyses:
   1. WER
   2. PER
   3. AFER

$$\frac{Insertion+Substitution+Deletion}{N}$$

RESULTS: WER

- Healthy speech > oral cancer speech (see Table 1)
  - Mann Whitney U test: W=0, p=0.0043 → **significant**

Table 1. Overview of the word recognition errors in percentages. **Blue bold** numbers indicate for which participant in each speaker group the ASR system achieved the best performance per column. **Orange bold** numbers represent the worst ASR performance per column for both speaker groups.

	Correct	Substitutions	Deletions	Insertions	WER
Healthy (n = 5)					
01	87.8	10.9	**1.2**	**2.8**	15
07	87.1	11.6	**1.3**	3.7	16.7
08	85.9	12.8	**1.3**	3.9	18
09	**84.7**	**14**	**1.3**	**7.1**	**22.4**
12	**89.5**	**9.3**	**1.2**	4.4	**14.9**
Mean	*87*	*11.7*	*1.3*	*4.4*	*17.4*
SD	*1.8*	*1.8*	*0.1*	*1.6*	*3.1*
Patient (n = 6)					
02	33.8	**62.7**	3.5	**27.4**	**93.6**
03	66.3	30.5	3.3	11.9	45.6
04	44	51.9	4.1	14.4	70.4
05	70.3	27	2.7	**8.2**	**37.9**
06	**72.7**	**26.1**	**1.2**	13	40.3
11	**32.8**	59	**8.2**	18.7	85.9
Mean	*53.3*	*42.9*	*3.8*	*15.6*	*62.3*
SD	*18.6*	*16.9*	*2.4*	*6.7*	*24.3*
Overall (n = 11)					
Mean	*68.6*	*28.7*	*2.7*	*10.5*	*41.9*
SD	*22*	*20.2*	*2.1*	*7.6*	*29.2*

- Patients with a glossectomy > patients with a mandibulectomy (see Figure 1)
  - Independent Samples t-test: $t(4)=1.03$, $p=0.36$, 95% CI [-34.44, 75.18]
    → failed to reach significance

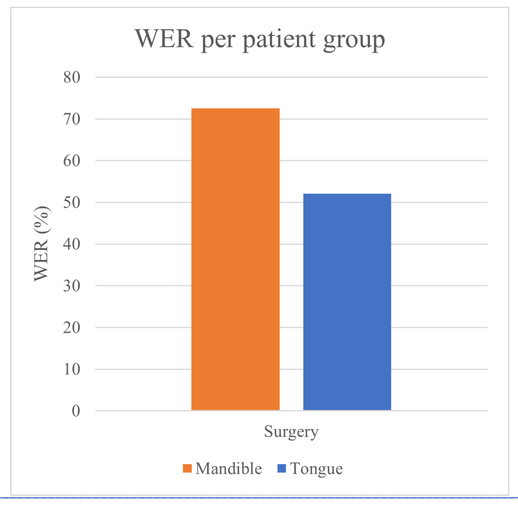

Figure 1. WER (%) results for oral cancer patients grouped by surgical treatment.

RESULTS: PER

- Threshold healthy speech: 10%
  - Phonemes with PER over threshold are /i(:), y, ɛ, ŋ, h, j/ (see Figure 2)
- Threshold oral cancer speech: 45%
  - Phonemes with PER over threshold are /i(:), k, ŋ, j/ (see Figure 2)

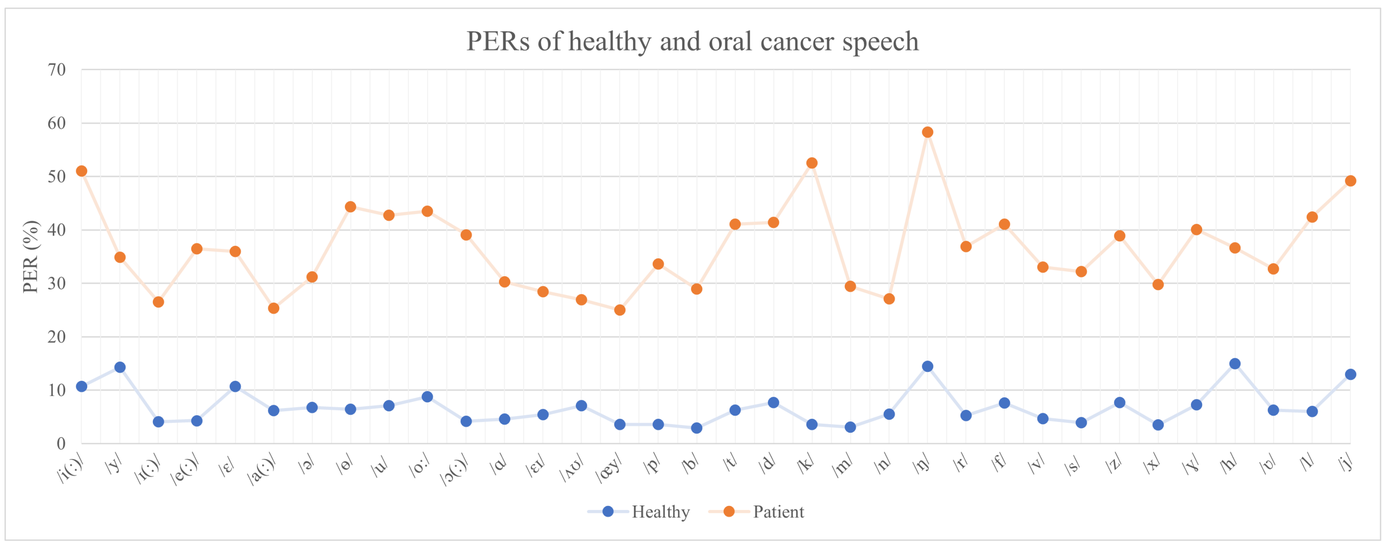

Figure 2. PER (%) results for healthy speakers and oral cancer patients.

- Generally, patients with a glossectomy > patients with a mandibulectomy
  - Except for /i(:), k, ŋ/ (see Figure 3)

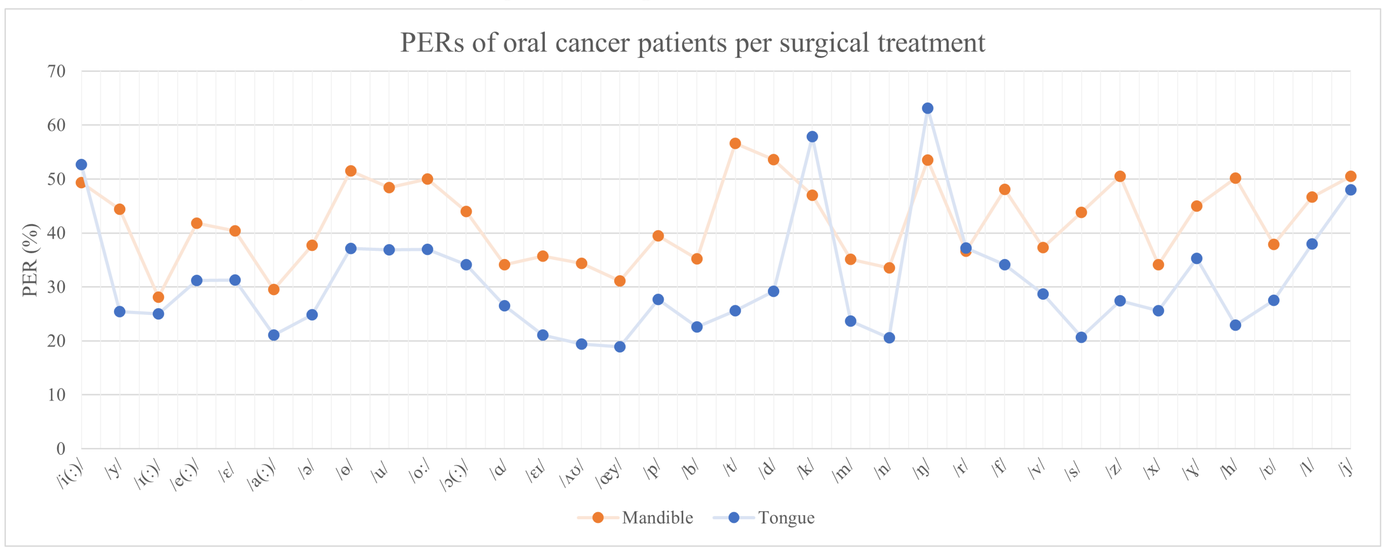

Figure 3. PER (%) results for oral cancer patients with mandible and tongue surgery.

DISCUSSION

- Particularly /k/ is challenging to capture for oral cancer speech, which is supported by AFER analysis, as plosives and velars elicit highest and second highest recognition error rates in oral cancer speech
- In contrast with our expectations, sibilants were relatively well captured
- Speech of patients who underwent a mandibulectomy obtains higher recognition error rates than the speech of patient who underwent a (partial) glossectomy, although the difference fails to reach significance
- Amount of data in this study was limited, and caution should be taken regarding the generality of our results
- Thus, plosives, especially /k/, elicit higher recognition error rates in Dutch oral cancer speech and the type of surgical treatment slightly affects ASR performance

REFERENCES

1. Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In International conference on machine learning, pages 1764–1772. PMLR.
2. Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., and Scharenborg, O. (2022). Low-resource automatic speech recognition and error analyses of oral cancer speech. Speech Communication.
3. Epstein, J. B., Emerton, S., Kolbinson, D. A., Le, N. D., Phillips, N., Stevenson-Moore, P., and Osoba, D. (1999). Quality of life and oral function following radiotherapy for head and neck cancer. Head Neck, 21(1):1–11.
4. Borggreven, P. A., Verdonck-de Leeuw, I., Langendijk, J. A., Doornaert, P., Koster, M. N., de Bree, R., and Leemans, C. R. (2005). Speech outcome after surgical treatment for oral and oropharyngeal cancer: A longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck, 27(9):785–793.
5. Laaksonen, J.-P., Rieger, J., Harris, J., and Seikaly, H. (2011). A longitudinal acoustic study of the effects of the radial forearm free flap reconstruction on sibilants produced by tongue cancer patients. Clinical Linguistics Phonetics, 25(4):253–264.
6. Matsui, Y., Ohno, K., Yamashita, Y., and Takahashi, K. (2007). Factors influencing postoperative speech function of tongue cancer patients following reconstruction with fasciocutaneous/myocutaneous flaps—a multicenter study. International Journal of Oral and Maxillofacial Surgery, 36(7):601–609.
7. Mooshammer, C., Hoole, P., and Geumann, A. (2007). Jaw and Order. Language and Speech, 50(2):145–176.
8. McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, pages 18–25.
9. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). ESPnet: End-to-End Speech Processing Toolkit. In INTERSPEECH, pages 2207–2211.
10. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.

Take a picture to **download** the **full paper**

university of groningen / campus fryslân