Low-resource automatic speech recognition and error analyses of oral cancer speech

Halpern, B.M.; Feng, S.; van Son, R.; van den Brekel, M.; Scharenborg, O.

DOI
10.1016/j.specom.2022.04.006

Publication date
2022

Document Version
Final published version

Published in
Speech Communication

Speech Communication 141 (2022) 14–27

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Low-resource automatic speech recognition and error analyses of oral cancer speech

Bence Mark Halpern [a,b,c,*,1], Siyuan Feng [b,1], Rob van Son [a,c], Michiel van den Brekel [a,c], Odette Scharenborg [b]

[a] Netherlands Cancer Institute, Amsterdam, The Netherlands
[b] Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[c] University of Amsterdam, ACLC, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Automatic speech recognition
Pathological speech
Low-resource
Oral cancer
Phoneme analysis

ABSTRACT

In this paper, we introduce a new corpus of oral cancer speech and present our study on the automatic recognition and analysis of oral cancer speech. A two-hour English oral cancer speech dataset is collected from YouTube. Formulated as a low-resource oral cancer ASR task, we investigate three acoustic modelling approaches that previously have worked well with low-resource scenarios using two different architectures; a hybrid architecture and a transformer-based end-to-end (E2E) model: (1) a retraining approach; (2) a speaker adaptation approach; and (3) a disentangled representation learning approach (only using the hybrid architecture). The approaches achieve a (1) 4.7% (hybrid) and 7.5% (E2E); (2) 7.7%; and (3) 2.0% absolute word error rate reduction, respectively, compared to a baseline system which is not trained on oral cancer speech. A detailed analysis of the speech recognition results shows that (1) plosives and certain vowels are the most difficult sounds to recognise in oral cancer speech — this problem is successfully alleviated by our proposed approaches; (3) however these sounds are also relatively poorly recognised in the case of healthy speech with the exception of/p/. (2) recognition performance of certain phonemes is strongly data-dependent; (4) In terms of the manner of articulation, E2E performs better with the exception of vowels — however, vowels have a large contribution to overall performance. As for the place of articulation, vowels, labiodentals, dentals and glottals are better captured by hybrid models, E2E is better on bilabial, alveolar, postalveolar, palatal and velar information. (5) Finally, our analysis provides some guidelines for selecting words that can be used as voice commands for ASR systems for oral cancer speakers.

1. Introduction

It is a great problem that many assistive technologies are only accessible to people with unimpaired speech. Often those who have the biggest need of such technologies are deprived of them. Oral cancer survivors are one such group of speakers. Approximately 500,000 people get diagnosed with oral cancer every year worldwide (Shield et al., 2017), of which 53,000 in the USA (The Oral Cancer Foundation, 2019) alone.

Oral cancer leads to speech impairments due to the (partial) removal of the tissues surrounding the tongue during surgery as part of the treatment of the oral cancer (Ward and van As-Brooks, 2014). Oral cancer speakers' speech impairments are predominantly on the articulatory level. Plosives (i.e. /k/, /g/, /b/, /p/, /t/, /d/) (Bressmann et al., 2009, 2004) and alveolar sibilants (i.e., /s/, /z/) (Laaksonen et al., 2011) have been found to be the most impacted (Halpern et al., 2020). In certain cases, patients are able to learn articulatory compensation techniques to adjust for the lost tongue tissue (Ward and van As-Brooks, 2014). Their impaired ability to speak affects their quality of life to a great extent (Epstein et al., 1999). This comes in addition to difficulty swallowing, chewing (Ward and van As-Brooks, 2014; Logemann et al., 1997), and decreased tongue mobility (Kappert et al., 2019) after operation.

This research focuses on building an automatic speech recognition (ASR) system for oral cancer speech. Such an ASR could have a large positive impact on survivors' quality of life and could be used in the objective evaluation of survivors' speech intelligibility during speech therapy (Windrich et al., 2008). To that end, this paper (1) presents a

* Corresponding author at: Netherlands Cancer Institute, Amsterdam, The Netherlands.
E-mail addresses: b.halpern@nki.nl (B.M. Halpern), s.feng@tudelft.nl (S. Feng), r.v.son@nki.nl (R. van Son), M.W.M.vandenBrekel@uva.nl (M. van den Brekel), o.e.scharenborg@tudelft.nl (O. Scharenborg).
[1] Authors contributed equally.

https://doi.org/10.1016/j.specom.2022.04.006
Received 1 March 2021; Received in revised form 13 February 2022; Accepted 22 April 2022
Available online 10 May 2022

newly collected database of English oral cancer speech; (2) investigates several approaches to building an ASR for oral cancer speech, where we specifically focus on the acoustic model to improve oral cancer speech recognition (and leave sophisticated language models and data augmentation for future research; see also the General Discussion); and (3) presents an analysis into the differences and similarities between oral cancer speech and normal speech.

Training a deep neural network (DNN) acoustic model (AM) for the automatic recognition of speech usually requires a large amount of labelled training data. In the case of oral cancer speech, though, we typically only have a very limited amount of labelled oral cancer speech data. This makes DNN AM training for oral cancer speech a low-resource problem. We investigate three hybrid approaches in low-resource ASR that previously have been shown to be competitive on low-resource tasks: (1) a retraining approach (Xu et al., 2015), (2) a speaker adaptation approach (Heck et al., 2017), and (3) a disentangled representation learning approach (Hsu et al., 2017) in order to leverage non-pathological, normal speech resources in DNN AM training for building AMs for oral cancer speech. (4) Due to the recent success of end-to-end (E2E) architectures, we additionally perform DNN AM retraining with a Transformer-based ASR architecture.

The acoustic model retraining approach leverages an AM pretrained on a healthy speech corpus and retrains this AM with oral cancer speech data. This approach has shown to be effective in improving acoustic modelling for pathological speech (Christensen et al., 2013; Liu et al., 2017), including dysarthria (Yilmaz et al., 2017; Hermann and Doss, 2020), and aphasia (Qin et al., 2018) for hybrid models. An effective multi-stage acoustic modelling method for dysarthric speech was proposed in Yilmaz et al. (2017).

Transformer-based E2E models are known to perform well when exposed to a large amount of training data and for standard, general-purpose ASR tasks (Karita et al., 2019). There is, however, limited research in pathological ASR using a Transformer-based architecture, with the exception of Harvill et al. (2021) for dysarthric ASR. However, the Transformer-based model achieves worse WER performance (even with data augmentation) than the current state-of-the-art (Hermann and Doss, 2020). The present study adopts a similar method to Yilmaz et al. (2017), and studies the efficacy of the retraining approach for the recognition of oral cancer speech using a hybrid and an E2E model.

The goal of speaker adaptation, or speaker adaptive training (SAT), is to normalise speaker variation contained in speech (Anastasakos et al., 1996), and is widely applied in general-purpose ASR systems (Gupta et al., 2014; Anastasakos et al., 1997; Miao et al., 2015; Cui et al., 2017). It is expected that speaker adaptation is even more important in oral cancer ASR, as oral cancer speech is much more variable than normal speech. We propose to use speaker adaptation, and particularly feature-space maximum likelihood linear regression (fMLLR) (Gales, 1998) based speaker adaptation, to suppress pathological speech sound characteristics in oral cancer speech, encouraging oral cancer speech representations to be more similar to those of normal speech. fMLLR has previously been successfully applied to improve pathological speech recognition performance (Hahm et al., 2015; Liu et al., 2017; Bhat et al., 2016). The resulting AM is expected to perform better on the oral cancer ASR task than that without speaker adaptation.

Disentangled speech representation learning aims to separate phonetic and speaker information in the speech signal into two feature representations in an unsupervised manner (Hsu et al., 2017), i.e., without the need of labelled speech data. One of the two learned representations, the phonetically-discriminative representation, is expected to retain the linguistic content in the original speech signal while suppressing speaker-dependent information. Conversely, the other learned representation is expected to capture speaker-dependent information and carry little phonetic information. The effectiveness of disentangled representation learning has been demonstrated for low-resource ASR (Feng and Lee, 2019; Feng et al., 2019) and noise robust ASR (Hsu and Glass, 2018). In the present study, we propose to apply this approach to suppress pathological speech sound characteristics while retaining the linguistic content in the oral cancer speech. Specifically, we adopt the factorised hierarchical variational auto-encoder (FHVAE) (Hsu et al., 2017) to perform disentangled speech representation learning. The learned phonetically-discriminative feature representation is used as the input feature to train a DNN AM for the oral cancer ASR task.

We further carry out an extensive phoneme-level and articulatory-level analysis in Section 4.2. The goal of this analysis is five-fold:

- Firstly, we want to find out what phonemes and articulatory features of the oral cancer speech are the most difficult to capture for current ASR systems trained on typical speech. This will allow us to investigate whether these sounds are the sounds that are known to be impacted in oral cancer speech or if ASR systems have problems with other sounds or aspects of oral cancer speech.
- Secondly, we want to pinpoint which phonemes and articulatory features contribute most to improvements in the proposed ASR systems. The motivation for this analysis is to identify performance bottlenecks, which will guide the development of future ASR systems. It is especially important to pinpoint phoneme classes where adding more oral cancer speech data is not expected to help. We would like to see which phonemes are better recognised by E2E models/hybrid models in the case of oral cancer speech. End-to-end models became superior to hybrid models on many ASR tasks, therefore we hypothesise that for certain sounds end-to-end models will be better. Determining which ones are better are essential for choosing the appropriate architecture for future pathological speech studies.
- Thirdly, we would like to compare the errors that the ASR architectures make on healthy and oral cancer speech. The goal of this analysis is to pinpoint which phoneme classes are specific to oral cancer speech, and which phonemes seem to be problematic for both kind of speech.
- Fourthly, the outcomes of the analyses will be used to provide guidelines on the selection of the words used for voice commands or stimuli for ASR systems aimed at oral cancer speakers. For example, if a particular class of phonemes are better recognised by the proposed systems than other phonemes, a voice command consisting mostly of phonemes from that class of phonemes can be selected. Such an analysis could bear meaningful lessons when deploying these systems to voice assistant tools or when these are used for objective evaluation of oral cancer speech.

Finally, it is well known that background noise negatively affects the performance of ASR systems (Cui and Alwan, 2005). Our dataset was collected from YouTube, which left us with little control regarding the noise in the audio. Therefore, it would be useful to quantify the influence of noise on the ASR performance, and compare it to the influence of speech severity. In Section 3.6, we perform an analysis to compare the influence of noise and speech severity in our ASR systems.

2. Dataset

In our experiments, we will use two datasets: a new, publicly available dataset we have recently collected containing English oral cancer speech[2]; and the Wall Street Journal (WSJ) dataset (Paul and Baker, 1992) containing English (*non-pathological*) read speech to leverage as training data for our baseline system and as a starting point for training our low-resource scenario ASR systems.

[2] https://karkirowle.github.io/oral_cancer_corpus/.

2.1. Oral cancer speech dataset

We manually collected 2.25 h of audio data containing English oral cancer speech from 10 different speakers from YouTube. Presence of oral cancer speech was determined by the content of the video and the authors' (B.H., R.v.S, M.v.d.B) clinical experience with such speakers. The audio was then manually cut to exclude music, healthy speakers, non-American English speakers, unintelligible speech, and other factors which could negatively influence recognition of the oral cancer speech. The resulting corpus has been automatically cut into chunks of 10 second. The cuts do not necessarily occur at natural pauses. When we transcribed the utterances, we tried to account for this as much as possible.

Baseline transcriptions were generated using the *Baseline* ASR system used in this study, which consisted of a DNN AM and a tri-gram language model (LM; see Section 3.1.1). Subsequently, these automatic transcriptions were manually checked and corrected by one of the authors (B.H.).

Table 1 shows the number of recordings and the amount of speech in minutes for each of the recordings of each of the speakers, as well as the speakers' gender. Since the total amount of oral cancer speech data is rather limited and because the total amount of audio for each speaker is highly variable, we carried out 5-fold cross-validation rather than creating separate training and test sets. A completely random, blind shuffling of the speakers into the five separate training and test sets would lead to (1) high variance in the observed WERs due to the large differences in the amount of audio used for training and testing in each possible partition, (2) high gender imbalance, i.e., in a completely random shuffling, an all-male train and all-female test set could easily occur. Therefore, to create the five training-test set combinations, the train and test set speakers are selected so that (1) the total audio used for training is always around 100 min (1.7 h), and (2) the gender balance of the train/test set varies within acceptable ranges, so that the training set contains at least two speakers of the same gender; and at least one speaker of that same gender is present in the test set. As a large portion of the audio data comes from the speaker with ID id011 (see Table 1), this speaker is always kept in the training set. The partitions are shown in Table 1. The speakers are either assigned to the training set or the test set, there is no overlap. The amounts of audio data in hours, the total numbers of words in the transcriptions, and the total number of audio files in the training and test data separated per gender are listed in Table 2 for each partition separately.

Table 1
Details of the oral cancer speech dataset and its train-test partitioning design for 5-fold cross-validation. Blue means train, while red means test.

Wav id	Spk id	Minutes	Gender	Partition index				
				1	2	3	4	5
1	id001	1.6	Female	Test	Test	Train	Train	Train
3		3.3		Test	Test	Train	Train	Train
10	id003	17.5	Female	Train	Train	Train	Test	Train
21	id007	12.8	Female	Train	Train	Train	Train	Test
23	id008	6.2	Female	Train	Test	Test	Train	Train
24		15.0		Train	Test	Test	Train	Train
18	id005	6.1	Female	Test	Test	Test	Train	Train
4	id011	1.4	Male	Train	Train	Train	Train	Train
5		4.2		Train	Train	Train	Train	Train
6		2.9		Train	Train	Train	Train	Train
7		3.2		Train	Train	Train	Train	Train
13		4.1		Train	Train	Train	Train	Train
22		11.9		Train	Train	Train	Train	Train
28		13.9		Train	Train	Train	Train	Train
26	id011/id009	13.3	Mixed	Train	Train	Train	Train	Train
30	id014	0.4	Male	Test	Test	Test	Train	Train
33	id016	1.8	Male	Test	Test	Test	Test	Train
34	id017	15.5	Male	Test	Train	Train	Test	Test

Table 2
Statistics of the training and test data in the 5-fold cross-validation scheme.

Partition index		1	2	3	4	5
Training set	Hours	1.77	1.68	1.76	1.67	1.78
	#words	17.2k	16.7k	17.3k	17.2k	17.5k
	#male audio files	7	8	8	8	9
	#female audio files	4	2	4	6	6
	#mixed audio files	1	1	1	1	1
Test set	Hours	0.48	0.57	0.49	0.58	0.47
	#words	4.7k	5.3k	4.6k	4.7k	4.4k
	#male audio files	3	5	3	1	1
	#female audio files	3	2	2	2	1
	#mixed audio files	0	0	0	0	0

2.2. Wall street journal corpus

The Wall Street Journal (WSJ) corpus is an American English read speech corpus (Paul and Baker, 1992). We used the `si284` set, which contains 37,416 speech utterances spoken by 283 speakers, for training. The total amount of data is 81.3 h. All speakers in the WSJ are healthy speakers.

3. Methods

The three approaches with the two different architectures to the automatic recognition of oral cancer speech will be compared against two *Baseline* ASR systems – one hybrid system and one E2E system – on the task of word recognition on the oral cancer speech test set. Word recognition performance is measured in word error rate (WER). We also report WER on the oral cancer speech training set, which is used in the analyses of the oral cancer recognition results (see Section 4.1.1). Fig. 1 present a schematic overview of the three approaches and the *Baseline* model implemented in the hybrid DNN-HMM architecture (top of Fig. 1). For ease of comparison of the three approaches, we used colours to indicate similarities (and differences) between the approaches: The blue colour indicates GMM-HMM training, the green colour indicates DNN AM (re-)training (the same approach is used for both architectures), and the orange colour indicates the feature representation method (only for the hybrid approach). The dashed boxes indicate the type of data that is used in the various stages of the pipelines of the three approaches. An overview of the training data, feature representations, and training methods of the three approaches and the *Baseline* models implemented in the hybrid and E2E architectures is provided in Table 3.

3.1. Baseline ASR systems

3.1.1. Baseline hybrid ASR

The *Baseline* hybrid ASR system is visualised in the left part of Fig. 1 (top) in the part of the pipeline that says "WSJ data", and consists of a hybrid DNN-hidden Markov model (DNN-HMM) AM only trained with WSJ. The input features of the *Baseline* system are 23-dimension filter banks (FBanks) appended by 3-dimension pitch features (Ghahremani et al., 2014). The 26-dimension features are further processed by contextual splicing $\{0, \pm1, \pm2, \pm3, \pm4, \pm5\}$ (following the recommendation in Kaldi[3]), i.e., each frame-level feature appended by its 5 left and 5 right frames, to capture longer temporal dependencies. This results in 286 $(26 \times (5 + 1 + 5))$ dimensions.

To obtain the phone labels for each speech frame of the WSJ data for DNN training, first a context-dependent GMM-HMM (CD-GMM-HMM) AM is trained from scratch with the WSJ training data and transcriptions using the standard Kaldi recipe (Povey et al., 2011). The

[3] wsj/s5/steps/nnet/pretrain_dbn.sh.

Table 3
Attributes of the hybrid and E2E models compared in this study. FB+P: FBank + Pitch feature. OC: oral cancer speech. '→': pretraining followed by retraining. '+': merging of the two datasets during training.

Architecture	Hybrid				E2E	
Method	Attributes					
	GMM-HMM		DNN		Transformer	
	Data	Input	Data	Input	Data	Input
Baseline	WSJ	MFCC	WSJ	FB+P	WSJ	FB+P
DNN AM/E2E ASR retraining	WSJ	MFCC	WSJ→OC	FB+P	WSJ→OC	FB+P
Baseline+OC	WSJ+OC	MFCC	WSJ+OC	FB+P	N/A	N/A
fMLLR for AM training	WSJ+OC	MFCC	WSJ+OC	fMLLR	N/A	N/A
FHVAE	WSJ+OC	MFCC	WSJ+OC	z_1	N/A	N/A

CMU dictionary[4] is used to map the words in the training data transcriptions to sequences of phonemes. The input features are 39-dimension MFCCs+Δ + $\Delta\Delta$. After CD-GMM-HMM AM training, the number of modelled HMM states is 3431. Frame labels are then obtained via forced alignment with the CD-GMM-HMM.

The DNN contains 5 feed-forward layers of dimension 1500 and a softmax output layer of dimension 3431 (equal to the number of HMM states). The DNN AM is trained using the WSJ frame labels as training labels and cross-entropy (CE) (Veselỳ et al., 2013) as the training criterion, and implemented based on Kaldi `nnet1`.[5] A 10% subset of training data is randomly selected for cross-validation (CV). The initial learning rate (LR) is 0.008, and is halved when no improvement of the loss value in the CV set is observed. Following the Kaldi `nnet1` convention, the training process is terminated if the LR is smaller than 1.5625×10^{-5}.

The *Baseline* ASR system uses a tri-gram LM trained with the transcriptions of the WSJ `si284` set. This LM is adopted consistently throughout all experiments in this paper.[6]

The *Baseline* ASR system achieves a WER of 6.7%[7] on the official WSJ test set `eval92`.

3.1.2. Baseline end-to-end (E2E) ASR

The *baseline E2E* ASR system adopts a transformer architecture (Karita et al., 2019) and is, like the Baseline hybrid model, only trained with the WSJ training material. The input features of the *E2E Baseline* system are 23-dimension FBanks appended by 3-dimension pitch features, the same input features as used for the *Baseline hybrid* system as described in Section 3.1.1. The transformer model parameters are taken mainly from the official ESPnet WSJ recipe[8]: 12 encoder layers and 6 decoder layers, all with 2048 dimensions; the attention dimension is 256 and the number of attention heads is 4; the convolution subsampling layer in the encoder has 2-layer CNN with 256 channels, stride with 2, and a kernel size of 3. The transformer model is trained with 50 epochs (no early-stopping), with a LR of 10.0, using a joint connectionist temporal classification (CTC)-attention objective (Kim et al., 2017) in which the CTC and attention weights are 0.3 and 0.7 respectively. Letters of the English alphabet are used as the basic subword units. The *E2E baseline* achieved 5.3% WER on the WSJ eval92 test set.

[4] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[5] The LF-MMI criterion (Povey et al., 2016) was found more effective than CE in dysarthric ASR (Hermann and Doss, 2020). However, our initial experiments using DNN AM trained with LF-MMI using the more recent `nnet3` in Kaldi showed no improvements over training using CE.

[6] RNNLM rescoring on top of tri-gram LM based results could lead to a WER reduction, however, this paper focuses on acoustic modelling, hence RNNLM rescoring to a hybrid model is not applied in this paper.

[7] This result falls short of state of the art (Zeghidour et al., 2018), mainly due to (1) the use of a tri-gram LM, and (2) the use of CMUdict without the extension to include the out-of-vocabulary words in the WSJ LM training data.

[8] `egs/wsj/asr1/conf/tuning/train_pytorch_transformer.yaml` from ESPNet.

3.2. Model retraining

In this approach, an ASR system is first trained with normal, i.e., in this case WSJ, speech data, and then retrained with oral cancer speech data. Sections 3.2.1 and 3.2.2 discuss the retraining approach applied to the hybrid and E2E ASR architectures, respectively.

3.2.1. Hybrid DNN AM retraining

The general framework of applying the retraining approach to a hybrid ASR system is illustrated in the top part of Fig. 1. The *Baseline* DNN AM described in Section 3.1.1 is chosen as the pretrained model and used as the starting point for retraining.

First, the *Baseline* DNN AM is used to force-align the oral cancer speech. Then, these alignments are used as labels to retrain the *Baseline* model. Preliminary experiments compared retraining some of the hidden layers vs. all hidden layers. The results showed that retraining all the hidden layers gave the best WER on the oral cancer speech test set. Therefore, *DNN AM retraining* in this study is always performed on all the hidden layers.

The loss function and stopping criterion of *DNN AM retraining* are the same as those for the *Baseline* DNN AM training. The initial LR was carefully tuned using the oral cancer speech data of partition 1 in the range of $\{0.002, 0.004, 0.008, 0.016\}$ because we discovered that with very limited amounts of oral cancer training speech data for *DNN AM retraining*, the WER performance on the oral cancer speech was sensitive to the initial LR.Our preliminary experiments showed that the optimal LR was 0.008, and it is used in all experiments in this paper.

3.2.2. E2E ASR retraining

The *baseline E2E ASR* system described in Section 3.1.2 is chosen as the pretrained model and used as the starting point for retraining. We carried out *E2E ASR retraining* on all the encoder and decoder network layers of the pretrained model, in order to be consistent with the setup in *hybrid DNN AM retraining* (see Section 3.2.1). Similarly, as in the case of *hybrid DNN AM retraining*, we experimentally found the performance of *E2E ASR retraining* is sensitive to the LR. Our results indicated that the optimal LR for transformer is 0.5, and it is used in all *E2E ASR retraining* experiments in this paper.

3.3. Speaker-adapted features for acoustic modelling

The idea of fMLLR is to map acoustic speech features from the original unadapted space to a speaker-adapted space, so that the adapted features are less dependent on speaker identities. This is realised by the estimation of speaker-specific transform matrices and bias vectors.

Mathematically, let $\boldsymbol{o}_t^s$ be an unadapted speech feature at frame t, spoken by speaker s. fMLLR estimates a matrix $\mathbf{A}^s$ and a bias vector $\boldsymbol{b}^s$, and transforms $\boldsymbol{o}_t^s$ to $\hat{\boldsymbol{o}}_t^s$ by,

$$\hat{\boldsymbol{o}}_t^s = \mathbf{A}^s \cdot \boldsymbol{o}_t^s + \boldsymbol{b}^s, \tag{1}$$

where $\hat{\boldsymbol{o}}_t^s$ is the corresponding speaker adapted feature. The estimation of $\mathbf{A}^s$ and $\boldsymbol{b}^s$ can be realised by an expectation-maximisation (EM)

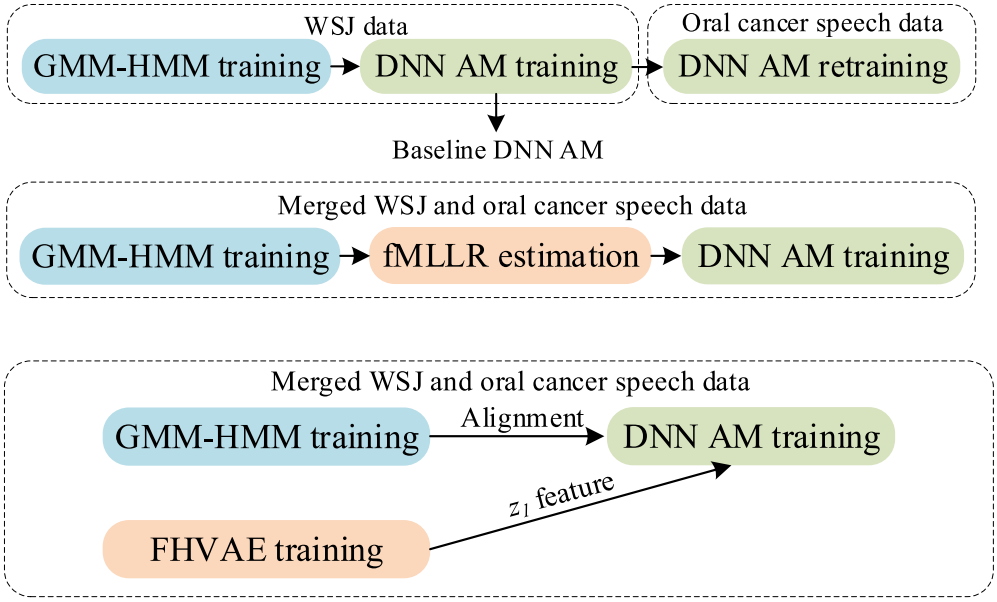

Fig. 1. (Top) Schematic overview of the *DNN AM retraining* approach. The left-most part, indicated with the dashed lines, shows the *Baseline* model. (Middle) Schematic overview of the *fMLLR* for AM training approach. (Bottom) Schematic overview of the disentangled speech representation learning for AM training approach. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

algorithm proposed in Gales (1998). The speaker adapted features $\hat{o}_t^s$ are also often referred to as fMLLR features.

The use of fMLLR features in acoustic modelling for oral cancer speech is illustrated in Fig. 1 (middle). The oral cancer speech data and WSJ data are merged to train a CD-GMM-HMM AM from scratch using the training procedure of the *Baseline* ASR system (see Section 3.1.1), except that here we also include the oral cancer speech data in the training of the CD-GMM-HMM AM model. Subsequently,fMLLR-based SAT is performed on the CD-GMM-HMM AM to estimate speaker-specific matrices and bias vectors. After SAT, a new CD-GMM-HMM AM with fMLLR features as input features is trained. This model is denoted as the CD-GMM-HMM-SAT. The dimension of fMLLR features is 40. The number of HMM states modelled by the CD-GMM-HMM-SAT model is 5080. Next, frame alignments are generated with CD-GMM-HMM-SAT for both the WSJ and the oral cancer speech data. These alignments and fMLLR features are used as training labels and input features, respectively, to train a DNN AM for oral cancer ASR.

In short, the DNN training procedure and architecture follow the settings of the *Baseline hybrid* DNN AM training, except: (1) Training data consists of both WSJ and oral cancer speech; (2) The softmax output layer dimension is 5080; (3) Input features to the DNN AM are fMLLR features, instead of FBank+pitch features. This method is denoted as *fMLLR for AM training* (or *fMLLR* for simplicity), and is only carried out for the hybrid architecture.

To explicitly measure the efficacy of fMLLR-based speaker adaptation, we trained another DNN AM, which takes 23-dimension FBanks appended by 3-dimension pitch features as input, instead of fMLLR features. Other training and model settings are the same as the system with *fMLLR for AM training*. This system is referred to as *Baseline+OC*, where OC stands for oral cancer speech.

3.4. Disentangled speech representation learning for acoustic modelling

Disentangled speech representation learning is based on the assumption that speaker characteristics vary less within an utterance than the linguistic content does, while linguistic content tends to have similar amounts of variation within and across utterances (Hsu et al., 2017). The FHVAE model (Hsu et al., 2017), which learns to factorise segment-level and sequence-level attributes of sequential data into different latent variables, is applied to disentangle phonetic (linguistic) and speaker information in the speech signal.

The FHVAE's encoder encodes input speech data into segment-level (expected to capture phonetic information) and sequence-level (expected to capture speaker information) latent variables separately, and the FHVAE's decoder reconstructs the original speech based on both the segment- and sequence-level latent variables (Hsu et al., 2017). Mathematically, let $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ denote the latent segment variable and the latent sequence variable, respectively. $\boldsymbol{\mu_2}$ is the sequence-dependent prior,[9] named as *s-vector*. θ and ϕ denote the parameters of the generation (decoder) and the inference (encoder) models of the FHVAEs, respectively. Let $D = \{\boldsymbol{X^i}\}_{i=1}^{M}$ denote a speech dataset with M sequences. Each $\boldsymbol{X^i}$ contains N^i speech segments $\{\boldsymbol{x}^{(i,n)}\}_{n=1}^{N^i}$, where $\boldsymbol{x}^{(i,n)}$ contains a number of consecutive frames.

The joint probability for the FHVAE decoder to generate $\boldsymbol{X}$ is formulated as,

$$p_\theta(\boldsymbol{\mu_2})\prod_{n=1}^{N} p_\theta(\boldsymbol{z_1^n})p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2})p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n},\boldsymbol{z_2^n}). \tag{2}$$

In the FHVAE, the exact posterior inference is intractable. The FHVAE introduces an inference model q_ϕ to approximate the intractable true posterior as,

$$q_\phi(\boldsymbol{\mu_2})\prod_{n=1}^{N} q_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})q_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n},\boldsymbol{z_2^n}). \tag{3}$$

Details of the formulation of Eqs. (2) and (3) are described in Section S1.1. The FHVAE model is trained by optimising a discriminative segmental variational lower bound (see Equation (S.4) in Section S1.1). To let $\boldsymbol{z_2}$ learn speaker-dependent feature representations, speech utterances of the same speaker in the training data are concatenated into a single sequence before training the FHVAE model. By this means, i, originally defined as the sequence index, becomes equal to the speaker index, and $\boldsymbol{z_1}$ learns a speaker-independent representation.

[9] Conceptually analogous to the i-vector in speaker recognition, one vector corresponding to a sequence.

The use of z_1 features in acoustic modelling for oral cancer speech is illustrated in Fig. 1 (bottom). The GMM-HMM training and the training data are exactly the same as for the fMLLR speaker adaptation method (see Section 3.3). The FHVAE model training is implemented using open-source software developed by Hsu et al. (2017). We used FHVAE parameters in Feng and Lee (2019) in our experiments: The encoder and decoder of the FHVAE are both 2-layer LSTMs with a layer dimension of 256. The dimensions of z_1 and z_2 are 32. The input features to the FHVAE are fixed-length (10 frames) speech segments. Each frame is represented by a 13-dimensional MFCC with cepstral mean normalisation at the speaker level. During the inference of the z_1 features, the FHVAE input segments are shifted by 1 frame, in order to match the length between speech frames and inferred z_1.

After FHVAE model training, the z_1 features of the WSJ and oral cancer speech are extracted and used as input features for the DNN AM training (see Table 3). This system is referred to as *FHVAE*. Compared with the **fMLLR for AM training** system, the only difference in the *FHVAE* system is the input representation to the DNN (z_1 versus fMLLR).

3.5. Phoneme and articulatory feature analysis

In this section, we describe the error analysis of our trained ASR systems. As a reminder, these analyses have five aims:

- (1) to investigate if the errors made by the ASRs are the same as the known articulation problems in oral cancer speech;
- (2) to find which sounds are poorly/well recognised in the proposed ASR system and to find out which sounds are better recognised with hybrid/E2E architectures
- (3) to compare the errors of the ASR models on healthy and oral cancer speech;
- (4) to provide input to the design of voice commands for ASR systems used by oral cancer speakers.

In our analyses, we will use the phoneme error rate (PER), and the articulatory feature error rate (AFER) as error measures. These metrics are similar to the word error rate (WER), except that they are calculated and interpreted at the level of phonemes and articulatory features (see Section 3.5.1). Confusion matrices of each model will be created and compared with one another to answer our research questions.

Specifically, for our first aim, we are going to look at the worst-performing phonemes and AFs of the *Baseline* system. This analysis assumes that the errors the ASR makes are based on the pronunciation mismatch between oral cancer and WSJ speakers.

For (2), we will investigate which phonemes are consistently misrecognised in the different approach and architecture combinations, and will compare them in terms of PER and WER. We will investigate whether the different approaches show problems with specific (groups of) phonemes by analysing whether the models have problems capturing particular articulatory feature information by looking at confusion matrices of AFs, or whether these systems' performances are mostly data dependent. We are going to further compare the differences between the best performing Hybrid and E2E techniques. This comparison will allow us to investigate which sounds are better handled by the E2E architectures, and which sounds are better with Hybrid.

For (3), we are going to compare the PER and AFER performances of the E2E and Hybrid Baseline models on the oral cancer and the WSJ test set. We will denote the WSJ test set experiments as *Hybrid on Healthy* and *E2E on Healthy*. The comparative analysis will allow us to investigate whether the same phonemes are found relatively difficult to the ASR systems.

For (4), we are going to compare the approaches in terms of PER and AFER, and we are interested in which phonemes are recognised well. Phonemes that are recognised well should be preferred in voice commands.

Table 4
PoA (columns) and MoA (rows) for each phoneme. **Abbreviations from left to right**: Bilabial, Labiodental, Dental, Alveolar, Postalveolar, Palatal, Velar, Glottal.

MoA	PoA							
	B	LD	D	A	P	PAL	V	G
Plosives	p, b			t,d			k,g	
Nasal	m			n			ng	
Fricative		f,v	th,dh	s,z		sh,zh		hh
Affricate						jh,ch		
Approximant	w			l	y	r		

The complete code for the analyses can be found online.[10]

3.5.1. Phoneme error rates and articulatory feature error rate

The PER is calculated as follows. First, the reference (ground truth) sentences and the sentences predicted by the ASR (hypothesis) are converted to phoneme sequences using the CMUdict.[11] The CMUdict contains the ARPABET phonemic transcription of 133,896 English words. Note that we do not take stress into account: Vowels with different stress markers are all treated as the same vowel. Second, the ground-truth phoneme sequence and the hypothesised phoneme sequence are aligned using the Levenshtein distance. We call these alignments Levenshtein alignments. Then, the PER is usually defined as:

$$\text{PER} = \frac{\text{insertion} + \text{substitution} + \text{deletion}}{N}, \tag{4}$$

where N is the total number of phonemes in the ground truth phoneme sequence. We also calculate the PER for each individual phoneme f in question as:

$$\text{PER}_\text{f} = \frac{\text{insertion}_\text{f} + \text{substitution}_\text{f} + \text{deletion}_\text{f}}{N_\text{f}}. \tag{5}$$

The AFER is calculated similarly to the PER, the main difference being that the aligned phoneme sequences are converted to place of articulation (PoA) and manner of articulation (MoA) feature sequences following Table 4 prior to the calculations of the error rates. The AFERs are also reported with respect to each individual articulatory feature, i.e., for the plosives,

$$\text{AFER}_\text{plosives} = \frac{\text{insertion}_\text{plosives} + \text{substitution}_\text{plosives} + \text{deletion}_\text{plosives}}{N_\text{plosives}}. \tag{6}$$

We report the mean and standard deviations of PER and AFER over all five test set partitions. In these analyses, we focus on those phonemes that have on average at least 100 occurrences ($N = 100$) in the ground truth, as we believe that 100 occurrences are the bare minimum to make meaningful conclusions. When $N \leq 100$, the results might be influenced too much by data scarcity.

3.5.2. Confusion matrices

Confusion matrices are used in the error analyses to investigate which articulatory feature classes are difficult for the ASRs to capture and which articulatory features are easily confused (modelling error). Using the Levenshtein alignments, we obtain an alignment of the ground truth phoneme sequences and the hypothesised phoneme sequences and create confusion matrices of the phoneme misrecognitions. In our description of the results, we group the phonemes by their AFs.

[10] https://github.com/karkirowle/relative_phoneme_analysis.

[11] In this method, we assume that any errors we observe at the phoneme level are due to the misrecognition of an individual phoneme (leading to a misrecognised word) rather than due to the misrecognition of a word which then would lead to the misrecognition of the phoneme. As we are using a large lexicon for training the ASR (see Section 3.1.1), we think this assumption is reasonable.

Since we are interested in the improvement or degradation of AFs in the trained systems compared to the *Baseline*, the Baseline confusion matrix will be separately shown in absolute terms. For the other systems' confusion matrices, the *Baseline* absolute performance will be subtracted.

3.6. Noise analysis

In this section, we describe our analysis which aims to quantify the influence of noise versus speech severity on the per-recording WER performance.

When quantifying the amount of noise in an audio file, usually the signal-to-noise ratio (SNR) is the figure of interest. Most existing SNR estimation methods are based on measuring the energy content of speech and non-speech regions in a signal. In the case of pathological speech, it has previously been shown that Parkinson's speech and whispered speech can negatively affect the SNR estimation (Poorjam et al., 2018). In other words, it is possible to obtain low SNR estimates in pathological voices even though there is no real background noise present in the recordings.

In order to avoid quantifying noise level by an SNR estimation algorithm that is heavily influenced by the severity of the pathological speech, we wanted to ensure that the correlation between the SNR and severity is low. In order to do that, first, the speech severity of each recording was quantified by an expert listener. To that end, an American English speech language pathologist (SLP) was asked to rate the severity of each recording on a 5-point Likert scale (1: very severe speech, 5: healthy speech) by listening to (at least one) 10 s segment of a recording. (Note that the 10 s segment constraint is based on constraints from an on-going study for which these ratings have been originally collected). The important consequence from the perspective of our analysis is that for some utterances the ratings have higher resolution. By resolution, we mean the step size of MOS during ratings, using one rating only 1-2-3-4-5 is obtainable (step size of 1), using two utterances it is possible to obtain 1-1.5-2-2.5-3-3.5-4-4.5-5 (step size of 0.5). This is because in the case of multiple ratings, we take the mean of the ratings.

Next, for the calculation of the SNR, the gold standard NIST algorithm is used. The NIST SNR is calculated as follows. First, a signal energy histogram is calculated by computing the root mean square (RMS) in dB over a 20 ms analysis window, with a time shift of 10 ms. Typically, this results in a bimodal histogram, one peak (left-most) corresponding to the noise level, and the other peak (right-most) corresponding to the signal level. A raised cosine function is fitted to the noise peak with a direct search algorithm (Hooke and Jeeves, 1961), with the objective to minimise the Chi-squared distance. The midpoint of the raised cosine function is labelled as the mean noise power level. The raised cosine curve is then subtracted from the complete RMS histogram to obtain a "noiseless" histogram with a single peak. Then, the peak corresponding to the 95th percentile is defined to be the speech level. Subtracting the noise level from the speech level, the signal to noise ratio is obtained.

Subsequently, Spearman's correlation was calculated between the severity scores and the SNR level ($r = 0.12$, $p \geq 0.5$). The obtained low correlation means that the severity scores and the SNR level are not correlated, therefore the SNR values seem to be independent of the influence of speech severity. This means that our SNR estimates can be reliably used to estimate noise in the recordings.

Finally, to assess the influence of noise on the WER, we did a Pearson's correlation of the per-recording WER (mean across all test partitions) with the SNR for each experiment (SNR-WER r). We perform this analysis for each approach and architecture combination in the paper to see if there are architecture-specific differences in the influence of noise. Furthermore, to assess the influence of speech severity on the WER, we performed a Spearman's correlation of the per-recording WER with the speech severity score (SLP-WER ρ).

4. Results and discussion

4.1. ASR results

In Section 4.1.1, we first discuss the experimental results of the first five systems listed in Table 5, all of which adopt a hybrid DNN-HMM ASR architecture. Next, in Section 4.1.2, we discuss the experimental results of the retraining approach applied to the hybrid versus E2E ASR architectures.

4.1.1. Hybrid ASR results on the training and test sets

The word error rates (% WER) on the oral cancer speech data achieved by the *Hybrid Baseline* ASR system, the *Hybrid Baseline+OC* system and the three proposed hybrid systems discussed in Sections 3.2.1, 3.3 and 3.4 are shown in the top rows in Table 5. For each system, the training and test WER results are listed for each of the five training-test data partitions separately (see Table 1 for details) and averaged over all partitions. The training WER results are calculated only on the oral cancer training data. **Bold** results indicate the best performance on a particular partition or on the average of all partitions.

Table 5 shows that the *Hybrid Baseline* system has the highest training and test WER results of all the systems on all the data partitions (excluding E2E systems). Considering that the *Hybrid Baseline* system achieved a WER of 6.7% on normal speech (see Section 3.1.1), the high WER results for the *Hybrid Baseline* system indicate a severe mismatch between oral cancer speech collected for this study and speech in the WSJ corpus. Although there are several differences between the WSJ and the oral cancer data set (including recording conditions and speaking style (read speech vs. spontaneous speech)). The primary cause of this deterioration is most likely the difference in type of speech, i.e., healthy versus oral cancer speech.

Table 5 shows that the *fMLLR* method achieved the best test WER results overall and on four out of the five data partitions (partition 2 is the exception). The hybrid *DNN AM retraining* method achieved an average absolute WER reduction of 34.0% on the training data and of 4.7% on the test data compared to the *Hybrid Baseline* system. The only difference between the AM retraining method and the *Hybrid Baseline* system is the use of a small amount (less than 2 hours, see Table 1) of oral cancer speech data during training in the *DNN AM retraining* system. These results show that such a small amount of speech material already helps to adapt the DNN AM from healthy speech to oral cancer speech and leads to an improvement in recognition performance.

The *fMLLR* system achieved the best performance on the oral cancer test data, achieving an average absolute WER reduction of 7.8% compared to the *Hybrid Baseline* system, and 3.1% compared to the hybrid *DNN AM retraining* system. Not only does the *fMLLR* system outperform the hybrid *DNN AM retraining* system overall on the test data, it also has a better performance on most of the test data partitions (except partition 2). These results suggest that the *fMLLR* approach is better than the DNN AM retraining approach, both in terms of the average WER performance and the per-partition WER performance.

The better performance of the *fMLLR* approach compared to the hybrid *DNN AM retraining* approach is in part due to the merging of the oral cancer speech data with the normal speech data during training, which allows the *fMLLR* model to leverage phonetic information from both healthy speech and oral cancer speech — unlike the *DNN AM retraining* approach which only has access to the oral cancer speech during the retraining phase. A further 4.5% absolute WER reduction on the test data is due to using the fMLLR features (as can be seen when comparing the *fMLLR* system with *Hybrid Baseline+OC*), which allows the model to leverage speaker diversity information.

Interestingly, the hybrid *DNN AM retraining* method achieves the best performance on the training data of all tested systems (excluding E2E systems), but performs worse than the *fMLLR* method on the test data. This finding is likely due to overfitting of the hybrid *DNN AM retraining* method on the small amount of oral cancer training data. At

Table 5
The word error rates (% WER) on the oral cancer speech on the different training-test partitions separately and averaged over all five partitions. **Bold**: best performance among the five systems. For the Baseline and Transformer E2E baseline systems, both training and test oral cancer speech data are unseen to the system, while for the remaining systems, the oral cancer speech training data is seen to the systems but not the oral cancer speech test data.

System	Partition index											
	1		2		3		4		5		Average	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Hybrid baseline	78.6	59.7	74.2	75.7	74.0	76.8	74.5	74.9	76.8	65.6	75.6	70.6
Hybrid DNN AM retraining	44.3	55.8	34.7	68.7	40.8	69.5	39.3	71.2	49.3	64.2	41.7	65.9
Hybrid Baseline+OC	53.6	55.8	49.8	74.7	47.5	73.2	47.5	70.9	51.2	62.0	49.9	67.3
fMLLR for AM training	49.0	**52.2**	49.7	69.4	47.4	68.7	46.2	**68.1**	48.9	**55.7**	48.2	**62.8**
FHVAE	50.3	58.0	48.7	73.1	47.0	73.5	46.5	72.6	48.1	65.2	48.1	68.5
E2E baseline	78.6	62.0	74.9	75.5	74.6	76.6	73.6	80.1	76.7	68.3	75.7	72.5
E2E ASR retraining	**23.1**	53.4	**21.8**	**66.0**	**22.3**	**66.4**	**23.8**	71.0	**24.3**	58.0	**23.1**	63.0

the retraining stage of the *DNN AM retraining* approach, the training data consists of oral cancer speech only. The hybrid DNN AM seems to overfit on the small amount of oral cancer speech training data, which then leads to a less well generalisation to unseen (test) oral cancer speech data. On the other hand, the *fMLLR* method merges the oral cancer speech and normal speech throughout the AM training procedure. In the *fMLLR* approach, during training, the AM is trained to perform well on both the WSJ data and the oral cancer data. This alleviates the overfitting problem, and consequently leads to a better generalisation to unseen oral cancer speech test data compared to the hybrid *DNN AM retraining* method.

The *FHVAE* method achieves better WER performance on the test data than the *Hybrid Baseline* system but worse than the other tested systems. It does achieve the second best WER performance on the training set among all the systems, after the hybrid *DNN AM retraining* method. Notably, the *FHVAE* method performs slightly better than the *Hybrid Baseline+OC* system on the training data, and slightly worse on the test data. The only difference between the *FHVAE* system and *Hybrid Baseline+OC* is the input feature representation to the DNN AM training: the *FHVAE* system uses $\boldsymbol{z}_1$ while the *Hybrid Baseline+OC* uses FBank with pitch features. The comparison between the two systems indicates FHVAE-based disentangled representation learning is effective in alleviating speaker-dependent characteristics in the training data in a limited but consistent manner on all the five partitions. However, it does not generalise well to unseen test data. A possible explanation is the small amount of available oral cancer speech data seen during FHVAE training. In Feng and Lee (2019), the effectiveness of FHVAE in a low-resource ASR task is shown to be sensitive to the amount of in-domain training data, and was shown to be very limited when there are only around 2 hours of training data available. To further explore the effect of FHVAE in the oral cancer ASR task, more (unlabelled, as FHVAE is unsupervised) audio recordings from oral cancer speakers should be used, which we leave for future study. However, due to the unlabelled nature of the data, this would be substantially easier to collect in large quantities.

4.1.2. Comparison of the hybrid and E2E ASR architectures in the AM retraining approach

The WERs (%) on the oral cancer speech data achieved by the two E2E ASR based systems, i.e., *E2E Baseline* and *E2E ASR retraining*, are shown in bottom rows in Table 5. Table 5 shows that the *E2E Baseline*'s performance is slightly worse than that of the *Hybrid Baseline* system on both the training and test sets.

Comparison of the two retraining based systems, i.e., *E2E ASR retraining* and the hybrid *DNN AM retraining*, shows that retraining is more effective in the E2E architecture than in the hybrid architecture for the oral cancer ASR task, at least with the current amount of oral cancer retraining data: The absolute WER reduction achieved by retraining is 9.5% for the E2E model, and is 4.7% for the hybrid model. Moreover, the *E2E ASR retraining* system achieves consistently better WER performances across all the partitions than the hybrid *DNN AM retraining* system. The significantly lower training set WERs achieved by the *E2E ASR retraining* indicates stronger modelling capability of the transformer E2E architecture than the hybrid architecture.

The *E2E ASR retraining* system achieves an average test data WER (63.0%) comparable to the best (*fMLLR for AM training*) system which adopts the hybrid ASR architecture (62.8%). Taking all results together, we can conclude that a transformer E2E ASR architecture achieves a WER for oral cancer ASR that approaches but does not outperform the speaker adaptation based hybrid DNN-HMM system.

4.2. Phoneme and articulatory feature error analysis

In this section, we present the key results of the error analysis. Each subsection will try to answer one of the five research questions outlined in Sections 1 and 3.5. All analyses have been carried out on the five oral cancer speech test set partitions separately and then averaged.

4.2.1. What phonemes are difficult for the baseline ASR systems?

In order to answer this question, we will first look at the phoneme level results, followed by the articulatory level results of the baseline models. Finally, we will compare our results with articulation problems known from the literature.

The phoneme level results are presented in Fig. 2. The y-axis indicates the PER at the phoneme level. The x-axis shows each of the phonemes in our data set, grouped by manner of articulation. Each line indicates a different system. Shaded regions denote the standard deviation for each model across the 5 folds. As can be seen in Fig. 2, most phonemes obtain a PER between 40%–60%. This indicates that the speech recognition task is challenging. Looking at the blue line (*Hybrid Baseline*), we can identify peaks corresponding to /g/, /aa/, /p/, /th/, /uw/. In the case of the *E2E Baseline*, the most difficult phonemes are /g/, /th/, /uw/, /aa/, /ey/. These are the most difficult phonemes for the baseline ASR systems to recognise. We can see that with the exception of /p/ and /ey/, the systems find the same phonemes difficult.

The AF level results are presented in Figs. 3 and 4 (top panels). In the case of the *Hybrid Baseline* MoA, affricates have the highest error, followed by plosives, approximants, nasals, fricatives and then vowels. For PoA, palatal sounds are the worst captured, followed by velars, postalveolars, bilabials, dentals, labiodentals, alveolars, glottals and, finally, vowels. In the case of the *E2E Baseline* MoA, we observe the same order as in the case of *Hybrid Baseline*. For *E2E Baseline* PoA, the palatals are the worst, followed by glottals, labiodentals, dentals, velars, postalveolars, bilabials, alveolars and vowels.

Previous research has already indicated that particularly plosives (Bressmann et al., 2009, 2004), sibilants (Laaksonen et al., 2011) and some vowels (/aa/, /ih/, /uw/) (Takatsu et al., 2017; Jacobi et al., 2013) are impacted by oral cancer. We can see that plosives have the second worst AFER, with two plosives (/g/ and /p/) having a PER of over 60%. As for sibilants (in our analysis: (post)alveolar fricatives), we observe that /s/ and /z/ are both comparatively well captured

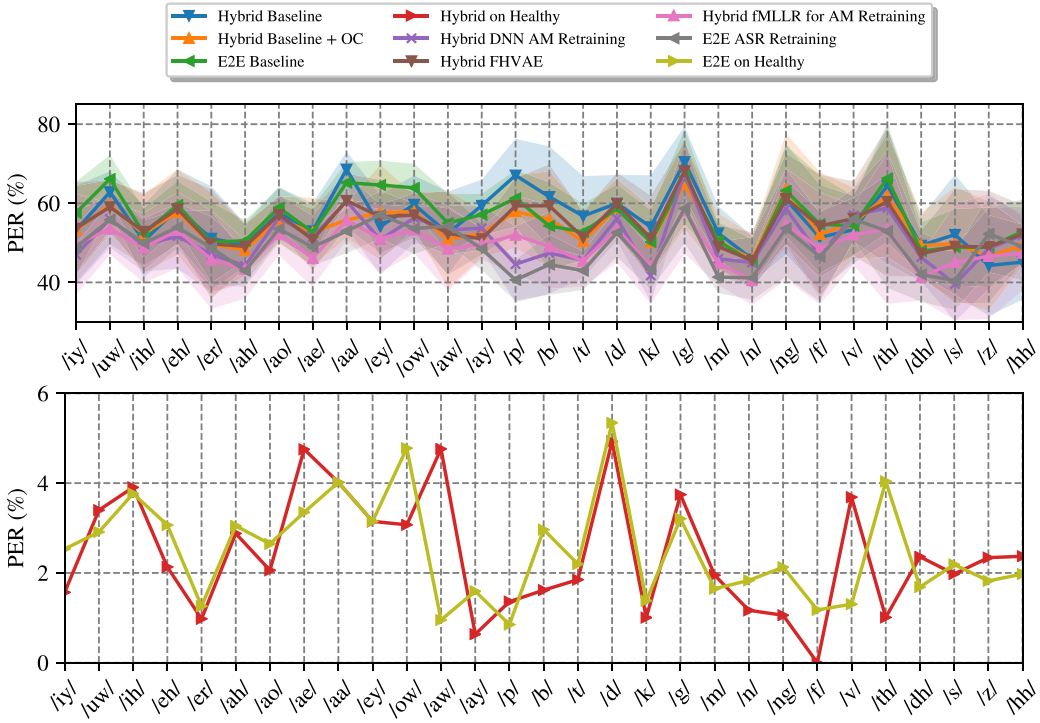

Fig. 2. Mean PER of each individual phoneme with $n \geq 100$. Shaded regions denote the standard deviation across the 5 folds. Line graph is used for ease of reading. Top panel describes PERs for the oral cancer dataset, while bottom panel describes PERs for the WSJ test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by the baseline ASR systems, showing that our systems did not have relatively more difficulty capturing sibilant information compared to other groups of phonemes. Finally, we can see that vowels are relatively well captured, with the exception of /aa/ and /uw/, which is consistent with the literature. The difficulty in recognising words with /ih/ as indicated by Takatsu et al. (2017) is not observed.

Overall, we see that those sounds that are known to cause articulatory problems after surgery for oral cancer speech, are also hard to recognise for the baseline ASR systems we tested. This is particularly the case for plosives and two vowels /aa/ and /uw/. In deviance to the literature, our systems did not have particular problems with sibilants. The reason for this difference is unclear: it might be that ASRs are more robust to variations in sibilant realisation. It would be interesting to confirm this with lisping speakers, where only sibilants are impacted. Interestingly, there were no sounds or articulatory features that were relatively hard for our ASR systems that were as yet unknown in the literature.

4.2.2. How well/poorly phonemes are recognised in the proposed ASR systems?

In order to investigate what techniques lead to a good recognition performance of oral cancer speech and what needs further investigation, we investigate which phonemes are improved and which ones are still misrecognised by analysing the produced error rates. Both for the phoneme and articulatory feature analysis, we additionally list the phonemes which seemed to work better with E2E architecture, and those which seemed to work better with Hybrid architecture.

As Fig. 2 shows, overall, the individual PER lies between 40% and 60%. A comparison of the different models shows that all the approaches generally improve the individual PERs compared to the baseline models (the blue line for the hybrid Baseline model and the green line for the E2E Baseline model), with a few exceptions, most notably the /hh/, /z/, /f/, /ey/ where particularly the *Hybrid Baseline*+OC orange line) and *FHVAE* (red line) models perform worse than the *Hybrid Baseline* model. In the case of the *E2E Baseline*, Hybrid Baseline+OC and FHVAE trained models perform worse on /b/. The hybrid systems outperform the E2E model in the case of /iy/, /uw/, /ih/, /eh/, /er/, /ao/, /ae/, /ey/, /ow/, /aw/ (therefore with most vowels), /k/, /g/, /v/, /dh/, /z/, and /hh/. The *E2E ASR retraining* system is better with /ah/, /aa/, /ay/, /p/, /b/, /t/, /d/ (therefore with most plosives), /m/, /n/ (all of the nasals), /g/, /f/, /th/, and /s/.

To further investigate whether certain (groups of) phonemes are consistently misrecognised, we investigate whether there is particular articulatory feature information that the models do not capture well. The extent to which the models can capture articulatory feature information is visualised in Figs. 3 and 4, the x-axis showing the different MoA/PoA and the number of phonemes (n) in each class, the y-axis showing the AFER. For PoA, palatal, postalveolar and velar sounds seem to be the most challenging, while for MoA these are affricates and approximants. Although all models in general improved the uptake of articulatory feature information (glottals being the exception), this was particularly the case for the Hybrid DNN AM retraining/E2E ASR retraining models for bilabial and plosive information. We observe that E2E better captures bilabial, alveolar, postalveolar, palatal, and velar information, while vowels, labiodentals, dentals and glottals are better captured by the hybrid models. For MoA (Fig. 4), the E2E model better captures plosive, nasal, fricative, affricate and approximant information, while vowel information is slightly better captured by the hybrid models, which actually has a larger impact on the overall performance (this can be observed by looking at the number of phonemes in each category, which is in parentheses).

We were interested if the difference between the AFER performances (i.e., vowels vs. affricates) was due to data scarcity in the phoneme classes to which the AFs were underlying. To investigate this, we performed a post-hoc Pearson's correlation analysis between the number of phonemes (of the AF class) (n) as the independent variable, and the PER performance as the dependent variable. The analysis found relatively strong effect sizes (*Hybrid DNN AM retraining*: 0.51, *fMLLR*: 0.55, *E2E ASR retraining*: 0.52, $p \leq 0.01$). Along with the fact that

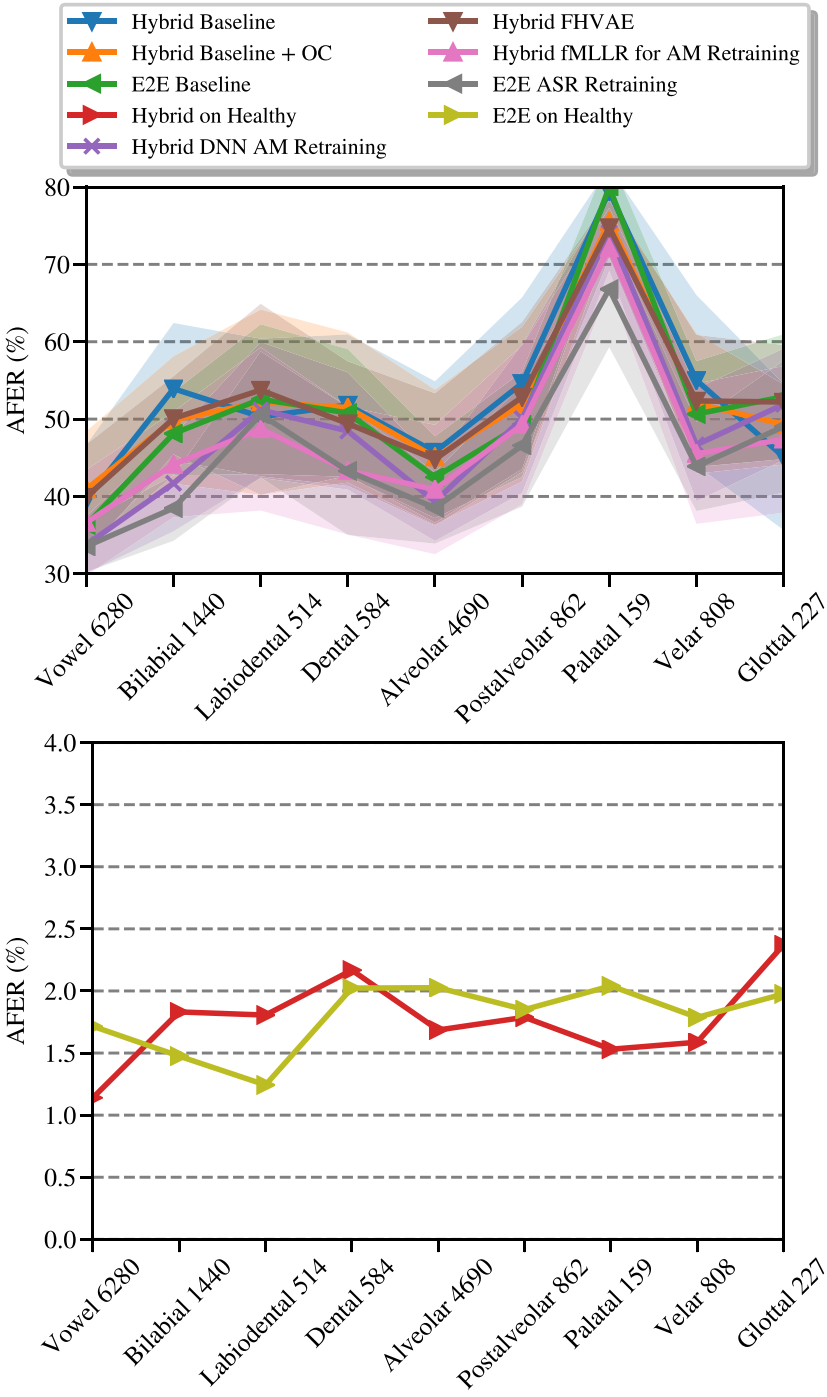

Fig. 3. **Top**: Comparison of AFER for PoA on the oral cancer test set. **Bottom**: Comparison of AFER for PoA on the WSJ test set. Mean N (phonemes in test set) rounded to three significant figures are in parentheses.

nearly all phonemes improve with our three approaches and for both architectures, we can conclude that the bottleneck of the performance seems to be mostly data-dependent. This means that it is important to collect corpora for oral cancer ASR in a phonetically balanced way, in order to have enough data to build good sound representations of each phoneme, including the rarer phonemes, such as glottals and palatals.

The confusion matrices in Fig. 5 enable further interpretation of these results. To better visualise the improvements, we have used relative confusion matrices for the proposed systems. In the case of relative confusion matrices, a green diagonal (more correct class) and a red off-diagonal (fewer incorrect classes) means improved classification. Also, note that for the insertion and deletion errors, a white line (meaning no errors) would be ideal for the absolute case, and a red or white line for the relative case (decreased errors or no change).

As a general remark, we can see that the majority of improvements in the *fMLLR* and *Hybrid DNN AM retraining* come from the reduction of deletion errors (red vertical line on the left side of the plot). For MoA, an additional part of this improvement comes from a reduction in substitutions of plosive sounds with fricative sounds compared to the *Hybrid Baseline*. Regarding PoA, we can see that (mainly) alveolar sounds and vowels were substituted with glottal sounds in the *Hybrid Baseline* model (light green vertical line in the middle), which is alleviated in the proposed approaches (red vertical lines in the middle of the plots). In the case of the *E2E ASR retraining model* we observe that

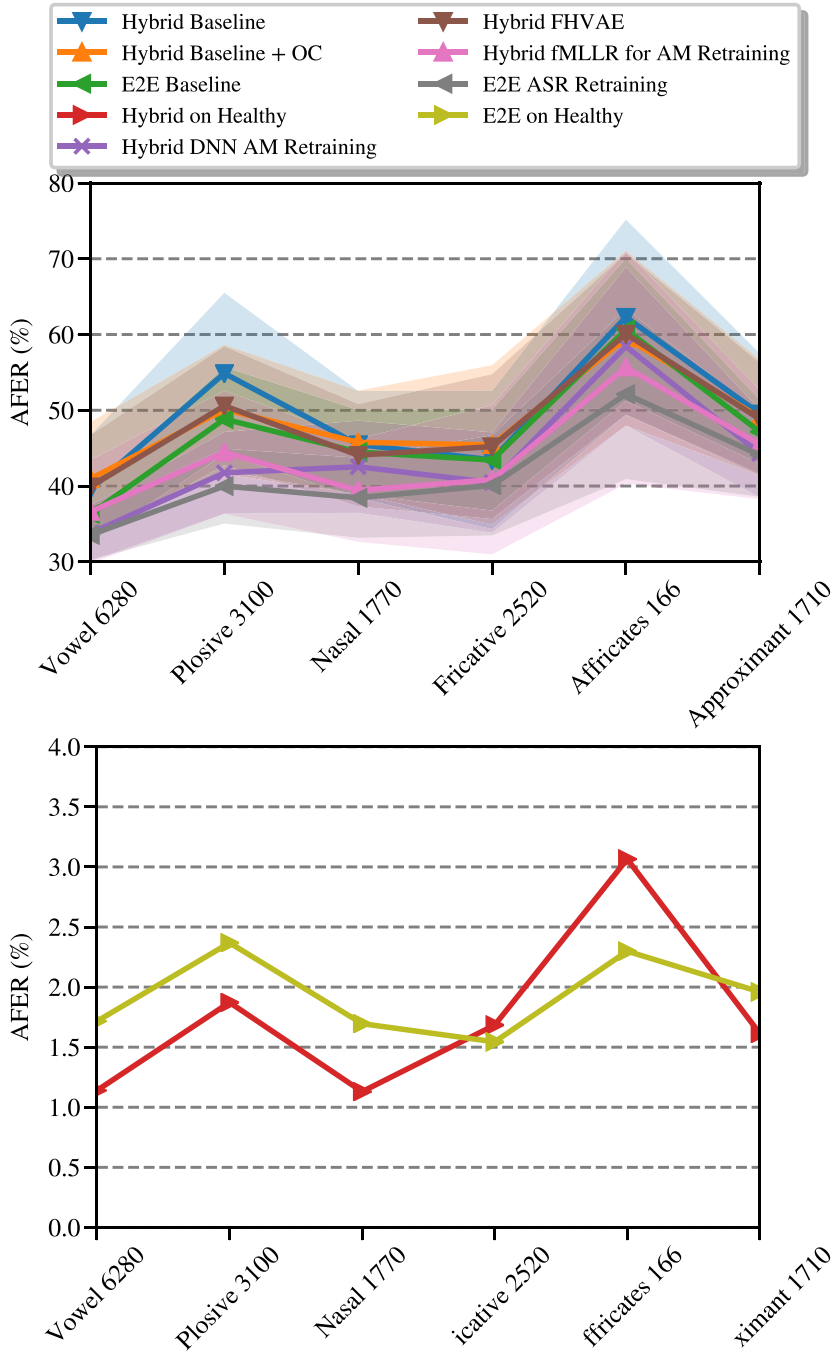

Fig. 4. Top: Comparison of AFER for MoA on the oral cancer test set. **Bottom**: Comparison of AFER for MoA on the WSJ test set. Mean N (phonemes in test set) rounded to three significant figures are in parentheses.

fewer sounds are classified as glottals, which makes the performance of the model worse on glottals overall compared to the *Hybrid DNN AM retraining*. Furthermore, a lot of phonemes are misclassified as dentals — it can be observed (vertical green lines) that the *E2E ASR retraining* model seems to make dentals as the "fallback" articulatory feature category.

We can summarise the findings as follows: (a) Plosive sounds are impacted in oral cancer speech, but speaker-adaptive training (*fMLLR* and *FHVAE*) and even a relatively small amount of training data (2 h; all proposed approaches) seem to alleviate these problems with the recognition of plosives. (b) Performance seems to be heavily data dependent, in general the number of phonemes is a good predictor of performance. (c) The "recognition" of /z/ and /hh/ is not improved over *Hybrid Baseline*, however this is partially explained by (b) as these two phoneme classes have relatively small amounts of training data (/z/ = 373 occurrences, /hh/ = 227 occurrences). This means that data augmentation techniques could be useful to alleviate the data scarcity problem. Overall, PER improvements brought by the proposed approaches compared to the baseline systems can be attributed to a general improvement in recognition performance across all phonemes. (d) In terms of manner of articulation, hybrid is only better compared to E2E on vowels — however, vowels have a large contribution to overall performance. As for place of articulation, vowels, labiodentals, dentals and glottals are better captured by hybrid models, while E2E better capture bilabial, alveolar, postalveolar, palatal and velar information.

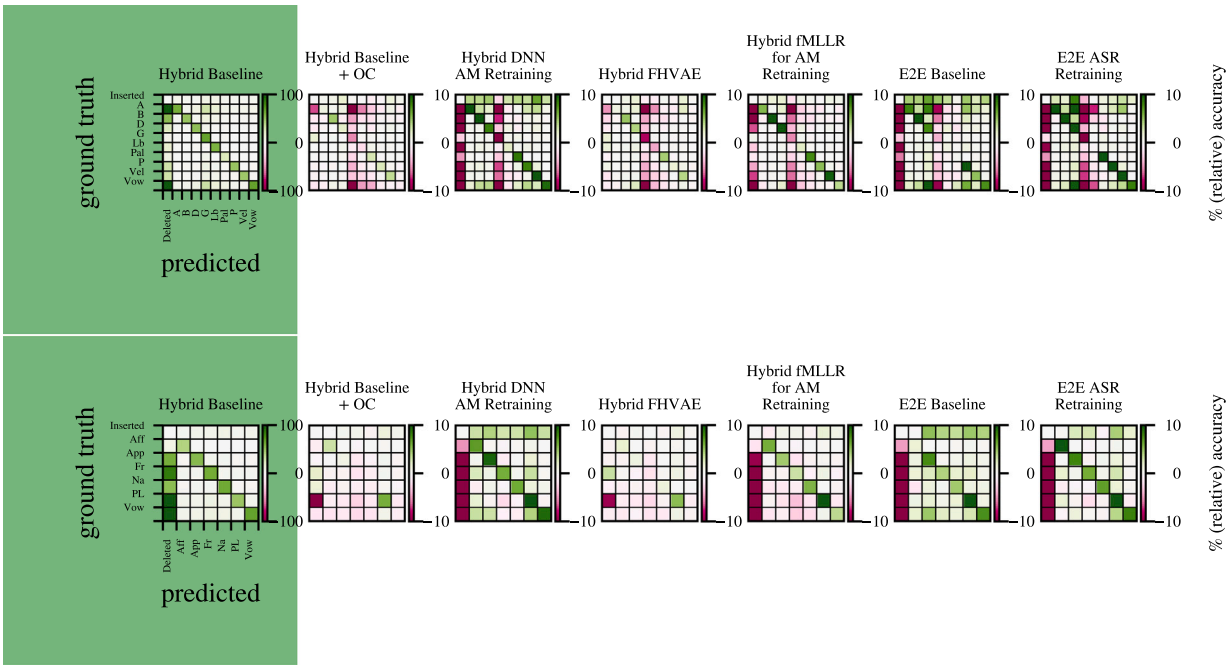

Fig. 5. Relative confusion matrices on PoA (top) and on MoA (bottom). Green diagonals and red off-diagonals mean better performance, while red diagonals and green off-diagonals mean worse performance. Green background denotes the absolute performances, while white background denotes relative performances. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Thus, in order to improve ASR for oral cancer speech, we conclude: (a) retraining approaches with even a small amount of extra training data can lead to substantial improvements for the AM; (b) Data augmentation techniques should be investigated for oral cancer ASR.

4.2.3. Do misrecognitions of oral cancer phonemes coincide with misrecognitions of healthy phonemes?

In this section, we would like to answer the question of whether the phoneme errors of the different approaches and architectures on oral cancer speech coincide with their errors on typical, healthy speech. In order to do that, we compare the PERs and the AFERs of the two baseline architectures (Hybrid and E2E Baseline) on both the oral cancer test set and the WSJ test set. (Note that this analysis is only carried out using the Baseline models as these are the only models that are only trained on healthy speech.)

The PERs on the oral cancer speech can be seen in the top panel of Fig. 2, while the PERs of the healthy speech can be seen in the bottom panel. We consider a phoneme relatively badly recognised in the case of oral cancer speech when the PER is over 60%. In the case of healthy speech, we set a threshold of 4%.

In the case of the hybrid architecture tested on healthy speech (*Hybrid on Healthy*) the phonemes /ae/, /aa/, /aw/ and /d/ are above the 4% threshold. In the case of the E2E architecture tested on healthy speech (*E2E on Healthy*), /aa/, /ow/, /d/ and /th/ are above the 4% threshold. For the hybrid architecture tested on oral cancer speech (*Hybrid Baseline*), the phonemes /uw/, /aa/, /p/, /b/, /g/, /ng/, /th/ are above the 60% threshold. For the E2E architecture tested on oral cancer speech (*E2E Baseline*), /uw/, /aa/, /ey/, /ow/, /p/, /g/, /ng/, /th/ are relatively badly recognised.

We can observe the following from these results. (1) The phonemes /aa/ and /d/ are relatively difficult for all architectures, independent of the type of speech used. (2) The phonemes /uw/, /p/, /g/, /ng/ are relatively more difficult in the case of oral cancer speech than in healthy speech.

This last finding (2) is partially consistent with the literature results discussed in Section 4.2.1, with the exception of /ng/. The /uw, p, and g/ sounds probably have a different pronunciation in oral cancer speech compared to healthy speech, leading to a worse recognition of these sounds by the Baseline models which have not been trained on oral cancer speech.

4.2.4. What voice commands should be used with oral cancer ASR?

When developing speech-driven systems or oral cancer speakers, it is preferable to base these on either the *Hybrid fMLLR* or *Hybrid DNN AM retraining* approaches as these are the two best systems. The results in the previous two subsections show that the phonemes that are best recognised by the DNN AM retraining are /s, k, ah, p, n/, while for *fMLLR retraining* these are /n, dh, ah, k, m/. So depending on which approach is used, we recommend selecting words containing these phonemes for the voice commands. Note that even though plosives are affected in oral cancer speech, our ASR results do not indicate that plosives should be excluded when designing voice commands.

4.3. How does noise in the dataset impact the results of the ASR systems?

Table 6 shows the influence of noise and speech severity on the WER. Each row corresponds to one audio recording with the corresponding WER rates on the different ASR models. From the low SNR-WER r correlation results, we can see that the impact of noise is generally low on the audio. The highest correlation between the SNR and the WER is for *E2E ASR retraining*, none of the SNR-WER correlations are significant. We can thus conclude that noise does not seem to have an influence on the WER results.

On the other hand, in all experimental conditions the speech severity seemed to be highly and significantly correlated with the WER results. The highest correlation is in the case of the *E2E Baseline*, followed by *Hybrid Baseline+OC*, *Hybrid DNN AM retraining*, *E2E ASR retraining*, and finally the *FHVAE* and the *fMLLR* methods. We can thus conclude that speech severity always has an influence on WER with the largest influence when there is no oral cancer data used for training, and the least influence when speaker-adaptive training is used.

Nevertheless, our subjective impression is that some recordings have quite challenging acoustic conditions for which speech enhancement techniques might be useful. We leave this for future research:

Table 6
Word error rates (%) and signal to noise ratios (SNR in dB) of the recordings in the dataset. Significance levels: * (p < 0.5), ** ($p < 0.01$), *** ($p < 0.001$).

Recording id	Baseline	Baseline + OC	DNN AM retraining	fMLLR	FHVAE	E2E baseline	E2E ASR retraining	SNR	SLP score
001	76.76	68.94	60.61	56.32	68.18	79.0	57.0	41.0	3.5
003	64.68	62.74	54.0	61.66	63.72	63.1	53.35	55.0	4.5
010	89.67	85.33	86.06	84.55	86.17	100.2	91.6	63.75	2.3
018	47.63	43.31	44.75	38.27	45.83	48.2	41.87	42.25	5.0
021	70.86	68.7	69.07	56.18	69.02	73.3	59.8	42.25	3.88
023	85.8	85.2	78.46	81.2	82.06	85.7	77.6	41.75	2.3
024	88.56	86.7	80.71	81.17	84.59	87.6	76.5	33.25	2.6
030	57.41	49.38	52.47	42.59	53.09	70.4	51.23	29.25	4.5
033	80.39	81.88	68.58	75.37	83.55	83.3	70.75	25.25	4.4
034	62.18	57.71	60.14	55.42	61.18	64.57	56.23	67.0	4.9
SNR-WER r	−0.04	−0.07	0.08	0.07	−0.05	−0.04	0.11	–	–
SLP-WER ρ	−0.93 ***	−0.91 ***	−0.91 ***	−0.87 **	−0.87 **	−0.93 ***	−0.91 ***	–	–

for instance, one approach could be for speakers who have multiple recordings (such as id008 and id011) to use a VoiceFilter-based speech enhancement (Wang et al., 2019). In that enhancement technique, an auxiliary recording is used to separate channel information pertaining to the speaker and background noise. Because there are many non-stationary noise sources in these audios, the VoiceFilter approach would probably be more beneficial than a spectral subtraction based approach, which is known to remove only stationary noise.

To summarise, we can conclude that speech severity impacts the WER performance to a great extent, and the impact of noise on the WER performance is substantially less.

4.4. Future work on the role of data augmentation

We hypothesise that some data augmentation techniques (such as pitch shift) would not work in the case of oral cancer speech, as the original speech is often already distorted beyond human comprehensibility. Existing literature for similar speech pathologies propose predominantly specific, custom techniques, i.e., the current state-of-the-art dysarthric ASR system uses speed perturbation (Hermann and Doss, 2020), other techniques propose voice conversion (Illa et al., 2021; Harvill et al., 2021). In the work of Harvill et al. (2021), it is also stated that data augmentation approaches seem to work better for high intelligibility pathological speakers. Therefore, we believe analysis of data augmentation techniques warrant a separate study, where effects such as the type of data augmentation, amount of data, and severity of speech can be separated in a controlled way.

5. Conclusion

In this paper, we presented a new dataset of American English oral cancer speech collected from YouTube. We investigated and compared two different DNN architectures on the task of oral cancer ASR with three different approaches: a *DNN AM retraining* (Hybrid, End-to-End) approach, an *fMLLR for AM training* approach, and an *FHVAE* approach. The *fMLLR* approach performed the best overall and achieved a WER of 62.8% on the oral cancer speech test set, which is a 7.8% absolute improvement over the *Hybrid Baseline*. Detailed error analyses on the recognition results of these approaches and architectures showed that (1) plosives and some vowels are challenging to recognise for the *Baseline* systems trained without oral cancer data, which is consistent with the literature on oral cancer speech which indicates that particularly plosives and some vowels are impacted by the removal of (parts of) the tongue due to oral cancer speech treatment. In contrast to the oral cancer literature, our models do not show the known problems with sibilants. In other words, we find that ASRs even without seeing oral cancer speech perform relatively well on sibilants of oral cancer speech. (2) The proposed approaches successfully alleviate the problems with the recognition of plosives and vowels. Furthermore, the proposed approaches and architectures do not show problems with particular phonemes, but rather their performance depends on the amount of training data for a given phoneme. Future research should therefore be directed towards data augmentation of particularly those phonemes with less training material, and speech enhancement techniques. (3) We find that it is mainly /uw/, /p/, /g/, /ng/ that are relatively difficult to recognise in the case of oral cancer speech, but not in the case of healthy speech (this analysis was only carried out on the Baseline systems). (4) For the development of voice command systems for oral cancer speakers, we propose to select words that include phonemes /s/, /k/, /ah/, /p/, /n/ for a system based on *Hybrid DNN AM retraining*, and /n/, /dh/, /ah/, /k/, /m/ for a system based on *fMLLR*. (5) A final analysis showed that channel noise in the recordings does not have an impact on the recognition performance of the models, rather the poor performance on the oral cancer speech is caused by the severity of the speech pathology.

CRediT authorship contribution statement

Bence Mark Halpern: Conceptualisation, Data Curation, Investigation, Formal Analysis, Methodology. **Siyuan Feng:** Methodology, Formal Analysis, Investigation. **Rob van Son:** Writing – review & editing, Funding acquisition, Supervision. **Michiel van den Brekel:** Writing – review & editing, Funding acquisition, Supervision. **Odette Scharenborg:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Noa Hannah (University of Illinois at Urbana-Champaign) for providing the severity ratings. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287 (TAPAS). The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.specom.2022.04.006.

References

Anastasakos, T., McDonough, J., Makhoul, J., 1997. Speaker adaptive training: A maximum likelihood approach to speaker normalization. In: ICASSP, Vol. 2. IEEE, pp. 1043–1046.

Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., 1996. A compact model for speaker-adaptive training. In: Proc. ICSLP, Vol. 2. pp. 1137–1140.

Bhat, C., Vachhani, B., Kopparapu, S.K., 2016. Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation. In: Interspeech. pp. 228–232.

Bressmann, T., Jacobs, H., Quintero, J., Irish, J.C., 2009. Speech outcomes for partial glossectomy surgery: Measures of speech articulation and listener perception indicateurs de la parole pour une glossectomie partielle: Mesures de l'articulation de la parole et de la perception des auditeurs. Head Neck Cancer 33 (4), 204.

Bressmann, T., Sader, R., Whitehill, T.L., Samman, N., 2004. Consonant intelligibility and tongue motility in patients with partial glossectomy. J. Oral Maxillofac. Surg. 62 (3), 298–303.

Christensen, H., Aniol, M.B., Bell, P., Green, P.D., Hain, T., King, S., Swietojanski, P., 2013. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In: Interspeech. pp. 3642–3645.

Cui, X., Alwan, A., 2005. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR. IEEE Trans. Speech Audio Process. 13 (6), 1161–1172.

Cui, X., Goel, V., Saon, G., 2017. Embedding-based speaker adaptive training of deep neural networks. In: Proc. Interspeech 2017. pp. 122–126. http://dx.doi.org/10.21437/Interspeech.2017-460.

Epstein, J.B., Emerton, S., Kolbinson, D.A., Le, N.D., Phillips, N., Stevenson-Moore, P., Osoba, D., 1999. Quality of life and oral function following radiotherapy for head and neck cancer. Head Neck http://dx.doi.org/10.1002/(SICI)1097-0347(199901)21:1<1::AID-HED1>3.0.CO;2-4.

Feng, S., Lee, T., 2019. Improving unsupervised subword modeling via disentangled speech representation learning and transformation. In: Proc. INTERSPEECH. pp. 281–285.

Feng, S., Lee, T., Peng, Z., 2019. Combining adversarial training and disentangled speech representation for robust zero-resource subword modeling. In: Proc. INTERSPEECH. pp. 1093–1097.

Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., 2014. A pitch extraction algorithm tuned for automatic speech recognition. In: ICASSP. IEEE, pp. 2494–2498.

Gupta, V., Kenny, P., Ouellet, P., Stafylakis, T., 2014. I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription. In: ICASSP. IEEE, pp. 6334–6338.

Hahm, S., Heitzman, D., Wang, J., 2015. Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization. In: Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. pp. 47–54.

Halpern, B.M., van Son, R., van den Brekel, M., Scharenborg, O., 2020. Detecting and analysing spontaneous oral cancer speech in the wild. In: Proc. Interspeech 2020. pp. 4826–4830. http://dx.doi.org/10.21437/Interspeech.2020-1598.

Harvill, J., Issa, D., Hasegawa-Johnson, M., Yoo, C., 2021. Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6428–6432.

Heck, M., Sakti, S., Nakamura, S., 2017. Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017. In: Proc. ASRU. IEEE, pp. 740–746.

Hermann, E., Doss, M.M., 2020. Dysarthric speech recognition with lattice-free MMI. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6109–6113.

Hooke, R., Jeeves, T.A., 1961. "Direct Search"solution of numerical and statistical problems. J. ACM 8 (2), 212–229.

Hsu, W., Glass, J.R., 2018. Extracting domain invariant features by unsupervised learning for robust automatic speech recognition. In: Proc. ICASSP. IEEE, pp. 5614–5618.

Hsu, W., Zhang, Y., Glass, J.R., 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In: Proc. NIPS. pp. 1878–1889.

Illa, M., Halpern, B.M., van Son, R., Moro-Velazquez, L., Scharenborg, O., 2021. Pathological voice adaptation with autoencoder-based voice conversion. In: Proc. 11th ISCA Speech Synthesis Workshop (SSW 11). pp. 19–24. http://dx.doi.org/10.21437/SSW.2021-4.

Jacobi, I., van Rossum, M.A., van der Molen, L., Hilgers, F.J., van den Brekel, M.W., 2013. Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. Ann. Otol. Rhinol. Laryngol. 122 (12), 754–762.

Kappert, K., van Alphen, M., Smeele, L., Balm, A., van der Heijden, F., 2019. Quantification of tongue mobility impairment using optical tracking in patients after receiving primary surgery or chemoradiation. PLoS One 14 (8).

Karita, S., Wang, X., Watanabe, S., Yoshimura, T., Zhang, W., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., 2019. A comparative study on transformer vs RNN in speech applications. In: Proc. ASRU. pp. 449–456.

Kim, S., Hori, T., Watanabe, S., 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: Proc. ICASSP. IEEE, pp. 4835–4839.

Laaksonen, J.-P., Rieger, J., Harris, J., Seikaly, H., 2011. A longitudinal acoustic study of the effects of the radial forearm free flap reconstruction on sibilants produced by tongue cancer patients. Clin. Linguist. Phon. 25 (4), 253–264.

Liu, Y., Lee, T., Ching, P.C., Law, T.K.T., Lee, K.Y.S., 2017. Acoustic assessment of disordered voice with continuous speech based on utterance-level ASR posterior features. In: Interspeech. pp. 2680–2684.

Logemann, J.A., Pauloski, B.R., Rademaker, A.W., Colangelo, L.A., 1997. Speech and swallowing rehabilitation for head and neck cancer patients. Oncology 11 (5).

Miao, Y., Zhang, H., Metze, F., 2015. Speaker adaptive training of deep neural network acoustic models using i-vectors. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (11), 1938–1949.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 357–362.

Poorjam, A.H., Little, M.A., Jensen, J.R., Christensen, M.G., 2018. A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 296–300.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proc. ASRU. pp. 1–4.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Proc. INTERSPEECH. pp. 2751–2755.

Qin, Y., Lee, T., Feng, S., Kong, A.P., 2018. Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning. In: Interspeech. pp. 3418–3422.

Shield, K.D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A.K., Bray, F., Soerjomataram, I., 2017. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA: Cancer J. Clin. 67 (1), 51–64.

Takatsu, J., Hanai, N., Suzuki, H., Yoshida, M., Tanaka, Y., Tanaka, S., Hasegawa, Y., Yamamoto, M., 2017. Phonologic and acoustic analysis of speech following glossectomy and the effect of rehabilitation on speech outcomes. J. Oral Maxillofac. Surg. 75 (7), 1530–1541.

The Oral Cancer Foundation, 2019. Oral Cancer Facts. The Oral Cancer Foundation, URL: https://oralcancerfoundation.org/facts/.

Veselỳ, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Interspeech. pp. 2345–2349.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J.R., Saurous, R.A., Weiss, R.J., Jia, Y., Moreno, I.L., 2019. VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In: Proc. Interspeech 2019. pp. 2728–2732. http://dx.doi.org/10.21437/Interspeech.2019-1101.

Ward, E.C., van As-Brooks, C.J., 2014. Head and Neck Cancer: Treatment, Rehabilitation, and Outcomes. Chapter 5: Speech and Swallowing Following Oral, Oropharyngeal, and Nasopharyngeal Cancer. Plural Publishing.

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U., Schuster, M., 2008. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr. Logop. 60 (3), 151–156.

Xu, H., Do, V.H., Xiao, X., Chng, E.S., 2015. A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition. In: Interspeech. pp. 2132–2136.

Yilmaz, E., Ganzeboom, M., Cucchiarini, C., Strik, H., 2017. Multi-stage DNN training for automatic recognition of dysarthric speech. In: Interspeech. pp. 2685–2689.

Zeghidour, N., Xu, Q., Liptchinsky, V., Usunier, N., Synnaeve, G., Collobert, R., 2018. Fully convolutional speech recognition. arXiv preprint arXiv:1812.06864.