Interspeech 2023 마비말 장애 관련 논문

최예린

서강대학교 인공지능학과

Email: lakahaga@u.sogang.ac.kr

2023.7.3

Interspeech 2023 – Dysarthric Speech Assessment

- Automatic assessments of dysarthric speech: the usability of acoustic-phonetic features
- Classification of multi-class vowels and fricatives from patients having Amyotrophic Lateral S clerosis with varied levels of dysarthria severity
- Parameter-efficient Dysarthric Speech Recognition Using Adapter Fusion and Householder Transformation
- Few-shot dysarthric speech recognition with text-to-speech data augmentation
- Latent Phrase Matching for Dysarthric Speech
- Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification

- 6편의 논문 (Oral Session)
  - 4개의 논문이 공개되어있음 (밑에서부터 4개)

Parameter-efficient Dysarthric Speech Recognition Using Adapter Fusion and Householder Transformation

- **Interspeech 2023**
- **KULeuven**
- **Paper**

Overview

- **Dysarthric ASR 에서의 challenge : data scarcity, vast diversity**
- **Adapter is better than Full-finetuning**

- **Speaker 마다, 심각도 마다 그 음성 특징이 다양하게 나타남**
- **Parameter efficiency와 speaker adaptation을 모두 해보자**

⇒ **Adaptor Fusion for target speaker adaptation**

- **정상인 음성으로 ASR 모델 학습 -> 각 speaker마다 하나의 adaptor를 훈련 -> target speaker (새로운 speaker)에 대해서 fusion layer로 knowledge transfer를 하여 성능 최적화**

Model structure

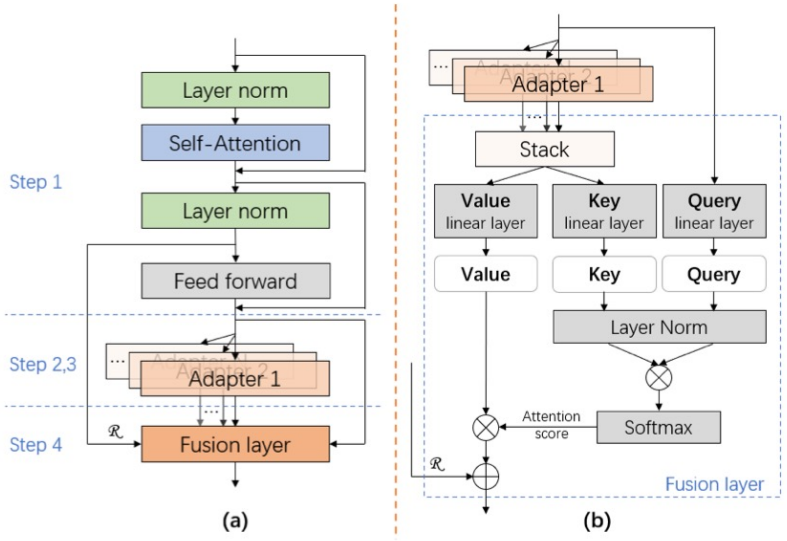


Figure 1: *(a) Transformer encoder layer with adapters and fusion layer, (b) details of the fusion layer.*

- **Base ASR Model**
  - **Transformer encoder decoder구조**
  - **Decoder는 CTC decoder에 대해서도 실험**
- **Adaptor와 fusion layer는 last encoder layer에만 추가**
- **Fusion layer**
  - **K,V 각 adaptor의 stacked output**
  - **Q: encoder의 output**
- **Speaker adaptation도 하지만, 각 speaker의 특징에서도 학습할 수 있도록 fusion layer를 추가**

훈련 방법

1. 정상인 음성에 대해 음성인식 모델 훈련
   - 300 시간의 Dutch Speech
2. 하나의 adaptor를 14명의 speaker에 대해서 훈련
   - Dysarthric Dutrch Speech (명령어 데이터)
3. 각 Source speaker에 대해 하나의 adaptor를 훈련
   - 이때 각 adaptor는 2번 adaptor의 weight에서 시작
4. Target speaker에 대해 fusion layer를 훈련

Result

Table 1: *Number of target-specific trainable parameters and CER in % when using at most 60 % of data.*

Name	#para	CER
Pretrain	-	49.98
FT-Enc	17.7M	1.62
FT-EncDec	27.2M	1.39
Pretrain-Adpt	-	13.27
Source-Adpt-avg	-	14.08
Target-Adpt	131.5k	4.40
Fusion-256dAtt+W	197.6k	**2.85**
Fusion-64dAtt+W	98.6k	**2.61**
Fusion-64dAtt	33.0k	8.38
Fusion-W	65.5k	3.03
Fusion-W_{UV}	-	6.62
Fusion-W_{Σ}	-	13.34
Fusion-P_{64}	32.8k	3.28
Fusion-W_{64}	33.0k	3.19
Fusion-64dAtt+W_{64}	66.1k	**2.79**

- **Target speaker에 대한 데이터는 60%만 학습**
- **Pretrain-Adapt**
  - **전체 speaker 함께 adaptor 훈련**
- **Target-Adapt**
  - **Pretrain-adapt를 target speaker에 대해 finetune**
- **Fusion + α**
  - **Fusion layer를 추가할 때 parameter를 줄여보고자 진행한 실험**
  - **K,Q layer의 dimension을 64로 설정할 때(줄일 때) 가장 성능이 좋았음**

Result

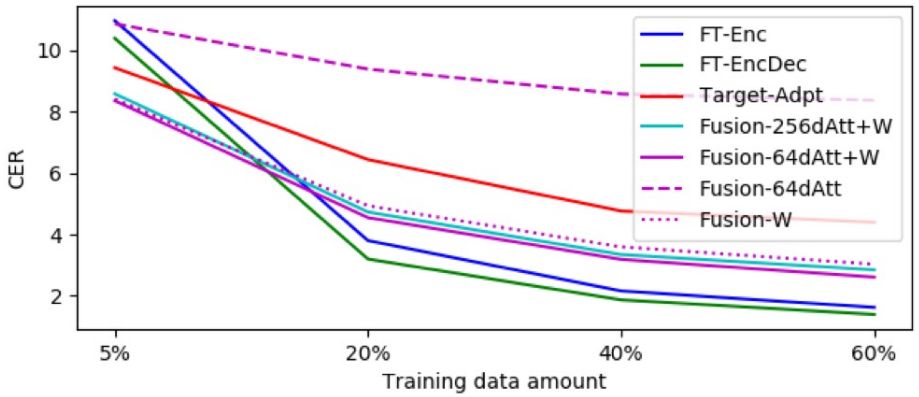


Figure 2: *CER results of different models when training on different amounts of target speaker data.*

- **Target speaker에 대한 훈련 데이터 양 조절**
- **Fusion layer에서 attention dimension을 줄였을 때가 성능이 full fine-tuning과 가장 비슷한 성능을 내었음**

Few-shot dysarthric speech recognition with text-to-speech data augmentation

- Interspeech 2023
- Idiap Research Institute
- Paper

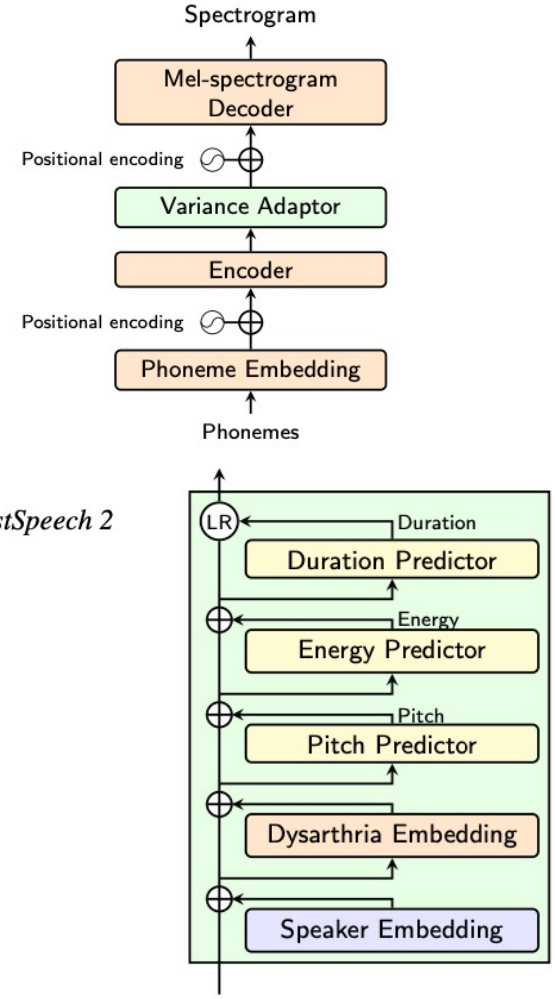
Overview

- 마비말 장애 음성의 ASR 성능 향상을 위한 데이터 증강 기법
- FastSpeech2를 이용하여 마비말 장애 음성 합성
  - 이때, variance adaptor에 dysarthric embedding을 추가
  - Speaker embedding을 VC system의 output을 활용

- 실험
  - TTS로 증강된 데이터를 사용해서 ASR 성능이 오르는지
  - 제안한 TTS 모델이 unseen speaker에 대해서도 데이터 증강이 가능한지?

- 결과
  - synthetic speech만으로 훈련하는 것은 한계가 있지만, 실제 음성과 합치면 실제 음성만 사용하는 것보다는 낫다
  - Unseen speaker에 대한 데이터 증강은 성능이 안 좋았음

Model Structure



(a) *FastSpeech 2*

(b) *Variance adaptor*

- **Modified FastSpeech2**
  - **Dysarthric embedding을 추가**
    - **speaker의 심각도에 대한 embedding**
  - **이때 심각도를 예측하지 않고, 주어진 값으로 사용**
    - **speaker에 dependent하기 때문에**
    - **우리와 같은 severity 체계 사용 (0, 1, 2)**
- **For unseen speaker**
  - **VC의 embedding을 speaker embedding으로 사용**
    - **X-vector나, id embedding을 쓰는 것 보다 좋았음**
    - **AdaIN-VC**

AdaIN-VC

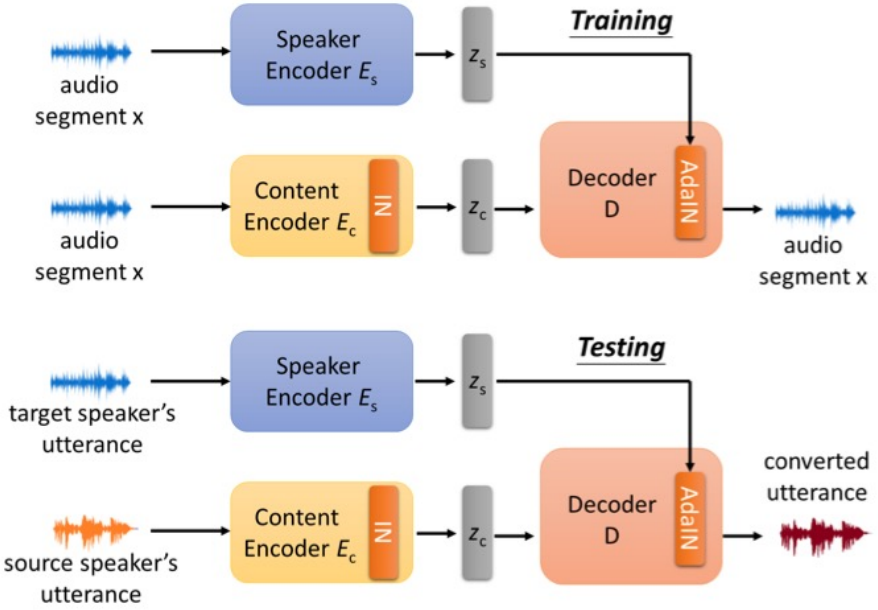
- **One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization**
  - **Interspeech 2019 / Hung-yi Lee, NTU / Paper, Github**

- **Instance Normalization을 사용**
  - ⇒ **Speaker 와 content 정보 분리 가능**
  - **Channel 단위로 normalization 적용**

$$\mu_c = \frac{1}{W}\sum_{w=1}^{W} M_c[w],$$

$$\sigma_c = \sqrt{\frac{1}{W}\sum_{w=1}^{W}(M_c[w]-\mu_c)^2 + \epsilon},$$

$$M'_c[w] = \frac{M_c[w]-\mu_c}{\sigma_c}$$



- **AdaIN**

$$M'_c[w] = \gamma_c \frac{M_c[w]-\mu_c}{\sigma_c} + \beta_c.$$

  - **γc, βc**
  - **Speaker encoder의 Output의 선형 변환**
  - **즉, speaker 정보**

증강 방법

- 대상 데이터셋: UA-Speech (단어 읽기 음성 set)
- 심각도 1,2 : 15명, 심각도 0 : 13 명
- 대상 텍스트 : block 1, 3
  - Block 2는 ASR 성능 평가에 활용 (block: 단어 set)

- UA-Speech의 훈련셋을 이용하여 TTS 모델 훈련하여 데이터 증강
  - 같은 데이터셋으로 AdaIN-VC 모델도 훈련
  - Vocoder (Hifi-GAN)도 훈련 (큰 성능 차이 없었음)

- Few-shot Speaker 실험을 위해 심각도 1,2에 대해 14명 훈련, 1명에 대해 평가
  - 구축한 TTS system이 unseen speaker의 5개 단어를 학습한 후, 합성음을 생성
  - 그 후, speaker dependent ASR을 훈련하여 ASR 성능 관찰

  ⇒ Unseen speaker에 대해서도 증강이 가능한지 실험

Result

Table 2: *Word error rates (WER) for each group of dysarthric speakers. For clarity, we also indicate whether the target speaker was seen during TTS training or not, where applicable.*

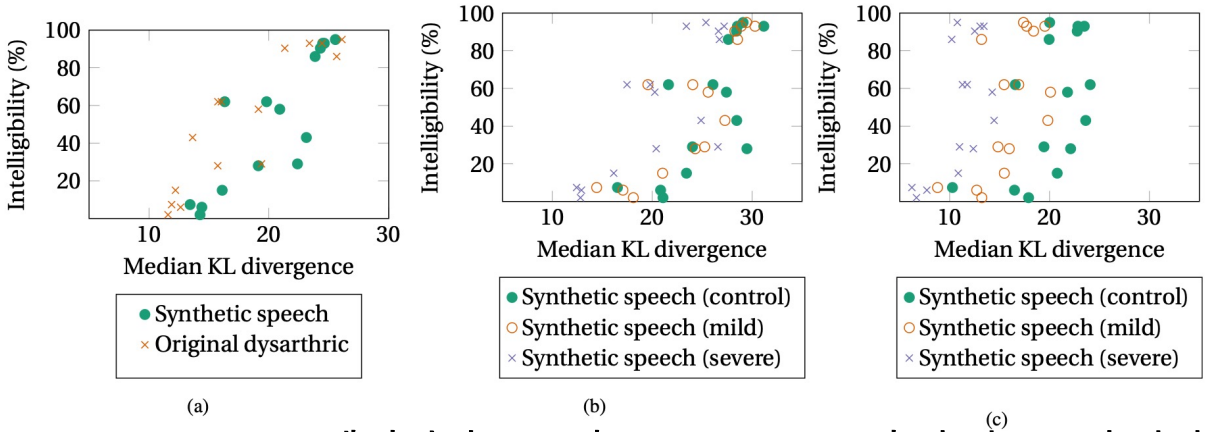
Systems	Seen	Sev.	Mod.-sev.	Mod.	Mild	Total
Baselines						
CTL	-	96.2	74.5	55.1	23.2	56.9
Top-line	-	70.3	42.7	38.2	24.0	41.3
+ CTL	-	65.8	34.3	25.3	15.4	32.8
Data augmentation						
TTS-aug	✓	70.8	38.7	33.6	18.5	37.6
TTS-aug4	✓	68.5	36.9	32.4	19.2	36.7
Few-shot						
F5-ctl	✗	99.6	99.1	98.1	92.0	96.5
+ CTL	✗	94.9	75.9	55.8	22.3	56.7
F100-ctl	✗	98.8	99.0	92.5	83.3	91.7
+ CTL	✗	93.8	75.6	51.7	21.8	55.4
F5-dys	✗	99.4	99.6	98.5	95.4	97.8
+ CTL	✗	94.4	76.1	53.9	22.5	56.8
F100-dys	✗	99.3	99.2	95.6	91.3	95.6
+ CTL	✗	94.5	72.7	52.4	20.6	54.6
F5-mix	✗	99.3	99.1	98.3	92.1	96.5
+ CTL	✗	94.7	75.8	55.6	21.4	56.3
F100-mix	✗	98.6	97.1	92.7	82.6	91.3
+ CTL	✗	93.7	72.7	50.7	20.9	54.2
TTS-only	✓	98.2	93.9	87.8	85.1	90.5
TTS-only4	✓	98.0	92.6	86.5	79.7	87.9

- 각 심각도별 Speaker-dependent ASR 성능 (WER)
- CTL : control, 즉, 정상인 음성
- Top-line: UA Speech 훈련 set 모두 학습 시킨 ASR 모델
- TTS-aug : UA Speech 훈련 set + 증강
  - Aug4: 합성음 수량을 4배로
- Few-shot
  - unseen speaker에 대해 합성한 후, ASR 학습
- ASR model
  - Acoustic: Kaldi-GMM with MFCC
  - Unigram decoder

- UA Speech+정상인 음성을 학습한 게 제일 좋음
- UA Speech + TTS : 증강 전보다 성능 향상
  - 증강 데이터 늘리면 더 향상
- Few-shot으로 생성한 데이터 : 성능 하락
- 증강 데이터만 학습 : 성능 더 안 좋음

합성된 데이터 분석

- 각 acoustic unit의 분포 간 KL divergence와 주관적 말 명료도의 상관도를 관찰
  - GMM 기반 ASR system에서 clustering 된 triphone 단위의 예측된 가우시안 분포
  - (a): 기존 speaker 정보 그대로 증강
  - (b): 하나의 utterance에 대해서 심각도 임베딩만 바꿔가면서 증강
    - A 에서는 기존 speaker 의 심각도를 따랐다면, 여기서는 한 speaker에 대해서 0,1,2 모두 증강해봄
  - (c) : few shot 증강



- (a): original 데이터의 unit 간 KL divergence와 말 명료도의 상관 계수 (r=0.90)와 증강된 데이터의 상관계수가 유사함(r=0.85)
  - 이는 a,b,c 모두에서 나타남
- 특히, b에서 심각도를 바꿔가면서 증강했을때, 각 utterance의 평균 duration이 심각도 별로 다름
  - 0: 1.2 초 / 1: 1.9 / 2: 2.6
  - 심각도 embedding이 length regulator에도 영향을 미치는 것을 확인

끝

감사합니다.