Automatic Dysarthric Speech Detection Exploiting Pairwise Distance-Based Convolutional Neural Networks

Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard

Idiap Research Institute

Virtual ICASSP

May 2021

FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

Outline

Dysarthric speech detection

- Dysarthria $\rightarrow$ disturbances of muscular control on speech production system
  - Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS)

Dysarthric speech detection

- Dysarthria → disturbances of muscular control on speech production system
  - Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS)

- Dysarthric speech detection: discriminating between normal and dysarthric speech

Dysarthric speech detection

- Dysarthria → disturbances of muscular control on speech production system
  - Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS)
- Dysarthric speech detection: discriminating between normal and dysarthric speech

- Subjective screening based on judgement of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis

- Automatic and objective detection method
  - Efficient, economical
  - Repeatable
  - Early diagnosis

Outline

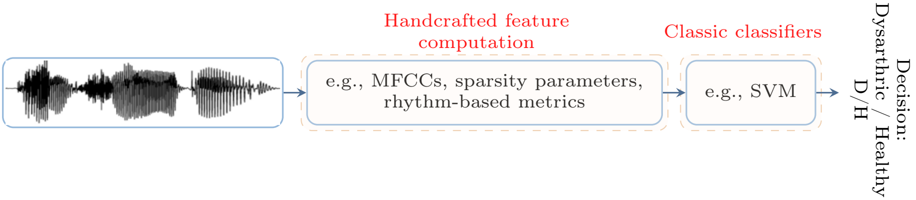
1. Automatic Dysarthric Speech Detection
2. State-of-the-art
3. Proposed Method
4. Experimental Results
5. Summary

State-of-the-art automatic dysarthric speech detection systems

- Focusing on connected speech analysis (words and sentences) → Crucial for assessment of dysarthria
  - Traditional machine learning approaches
  - Deep learning approaches

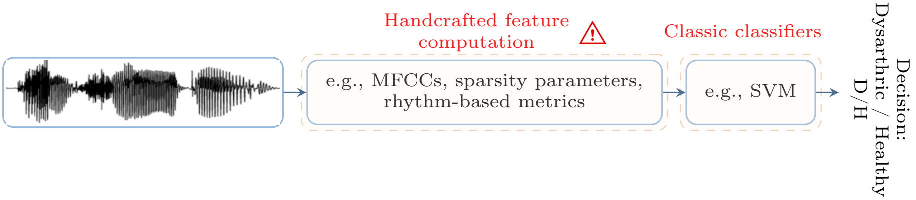
State-of-the-art automatic dysarthric speech detection systems

» Traditional machine learning approaches (Hegde et al., 2019; Kodrasi and Bourlard, 2020; Hernandez et al., 2020)



State-of-the-art automatic dysarthric speech detection systems

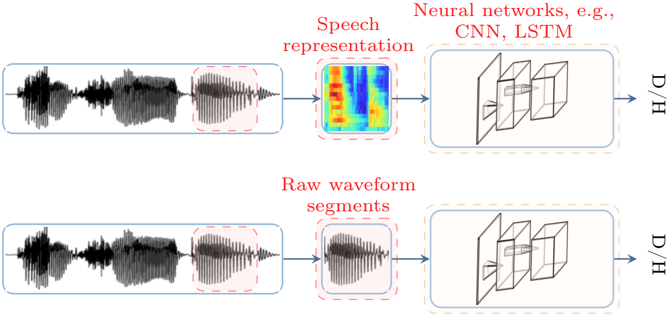
» Traditional machine learning approaches (Hegde et al., 2019; Kodrasi and Bourlard, 2020; Hernandez et al., 2020)



⚠ Abstract but important acoustic cues not characterized by handcrafted features

State-of-the-art automatic dysarthric speech detection systems

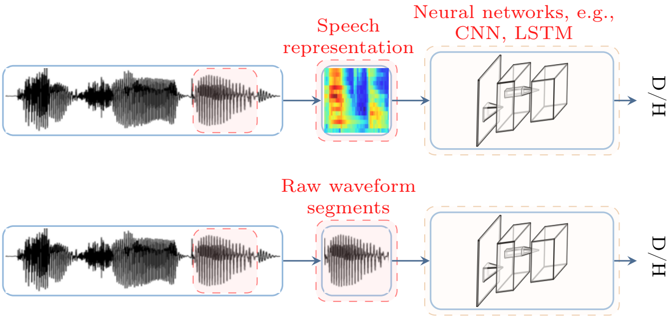
- Deep learning approaches → Data-driven approach using no prior knowledge
  - Exploit high-level abstract representations from low-level speech representations or raw waveforms
  - Challenge: alleviating overfitting associated with limited available dysarthric training data



State-of-the-art automatic dysarthric speech detection systems

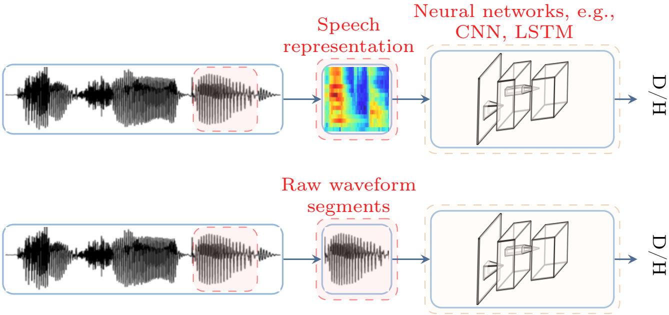
- Deep learning approaches → Data-driven approach using no prior knowledge
  - Challenge: alleviating overfitting associated with limited available dysarthric training data

Analysing (many) short segments of speech (Vasquez et al., 2017; Vaiciukynas et al., 2017; An et al., 2018; Mallela et al., 2020)



State-of-the-art automatic dysarthric speech detection systems

- Deep learning approaches → Data-driven approach using no prior knowledge
  - Challenge: alleviating overfitting associated with limited available dysarthric training data
  - ⚠ Analysing (many) short segments of speech → less robust to speaker variabilities (unrelated to dysarthria)



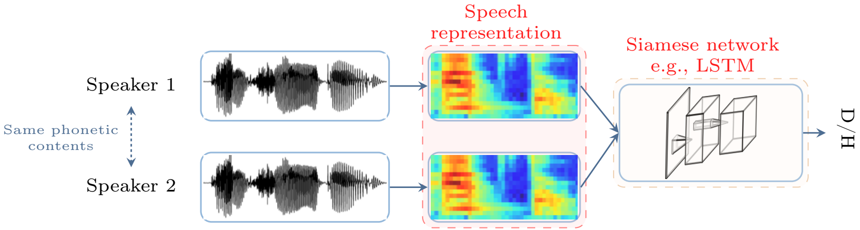
State-of-the-art automatic dysarthric speech detection systems

» Deep learning approaches → Data-driven approach using no prior knowledge

- Challenge: alleviating overfitting associated with limited available dysarthric training data

⚠ Analysing (many) short segments of speech

⚠ Training different LSTM Siamese networks for different utterances (Bhati et al., 2019)



Outline

CNN-based detection system exploiting pairwise distance matrices

- » Considering pairs of inputs representations; one of which is always from a healthy speaker (reference) and a test speaker
- » Computing frame-level distance matrices between reference representations and phonetically-mathced test representations → inputs to CNN for classification

Hypothesis

The frame-level distance matrix between two healthy utterances have different patterns (i.e., expected to be more quasi-diagonal) than between a healthy and a pathological utterances

CNN-based detection system exploiting pairwise distance matrices

- » Considering pairs of inputs representations; one of which is always from a healthy speaker (reference) and a test speaker
- » Computing frame-level distance matrices between reference representations and phonetically-mathced test representations → inputs to CNN for classification

Hypothesis

The frame-level distance matrix between two healthy utterances have different patterns (i.e., expected to be more quasi-diagonal) than between a healthy and a pathological utterances

✓ Pairwise training

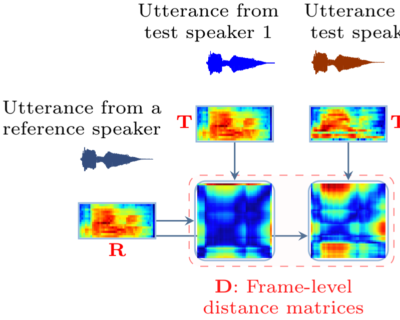
Advantageous for limited training data and for extracting features robust to unrelated speaker variabilities

✓ A single network can be used for different utterances

Since CNN operates on distance matrices

CNN-based detection system exploiting pairwise distance matrices

The frame-level distance matrix between two healthy utterances have different patterns (i.e., expected to be more quasi-diagonal) than between a healthy and a pathological utterances



Reference representation: $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_S]$
Test representation: $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_S]$
$\mathbf{r}_i$ and $\mathbf{t}_j$: reference and test feature vectors at time frame i and j.

Distance matrix $\mathbf{D}$ with (i, j)–th entry:
The distance d between $\mathbf{r}_i$ and $\mathbf{t}_j$

$$\mathbf{D}_{i,j} = d(\mathbf{t}_i, \mathbf{r}_j)$$

CNN-based detection system exploiting pairwise distance matrices

Input utterance representations

- Short-time Fourier transform (STFT) representation, e.g., in (Vasquez et al., 2017)
- Articulatory posteriors (APs)

CNN-based detection system exploiting pairwise distance matrices

Input utterance representations

» Short-time Fourier transform (STFT) representation

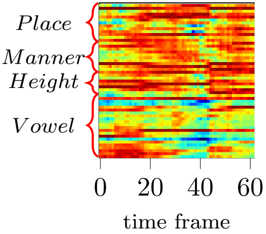
» Articulatory posteriors (APs) ✓

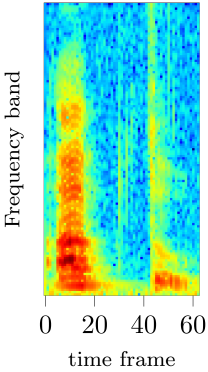
» APs $\rightarrow$ CNN-based phoneme-to-articulatory feature mapping trained using healthy speech data

- Mapping phone/phoneme into a set of multi-valued features based on the articulators used to produce it, e.g., manner, place, height, and vowel (Rasipuram and Magimai-Doss, 2016)
- Characterising articulation
- Robustness to noise
- Multilingual and crosslingual portability

CNN-based detection system exploiting pairwise distance matrices

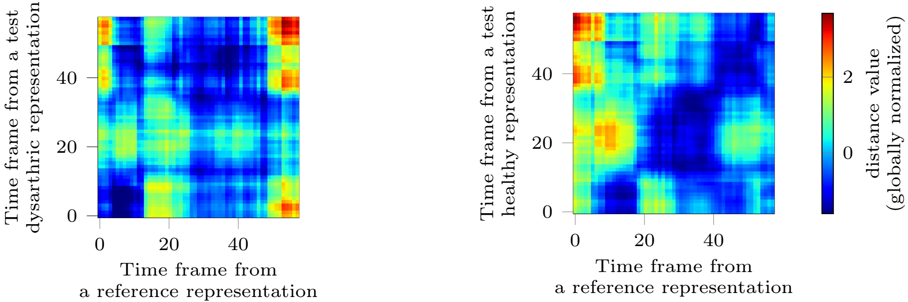
- AP and STFT representation of a sample utterance





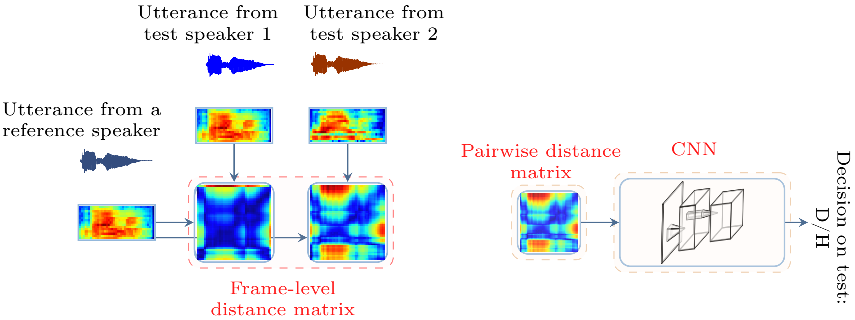
CNN-based detection system exploiting pairwise distance matrices

» Distance matrices computed from AP representations of a sample utterance from a pair of test dysarthric-reference and from a pair of test healthy-reference speakers.



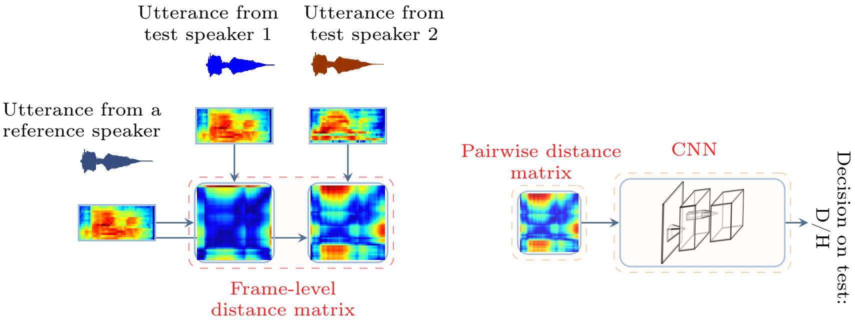
CNN-based detection system exploiting pairwise distance matrices

» Applying CNNs on frame-level distance matrices computed from user-defined representations of utterances



CNN-based detection system exploiting pairwise distance matrices

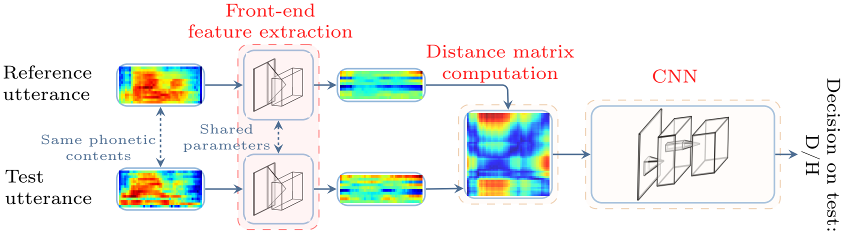
- Applying CNNs on frame-level distance matrices computed from user-defined representations of utterances
  - ⚠ The user-defined representations might not be optimal for healthy and dysarthric speech detection



CNN-based system exploiting pairwise distance matrices

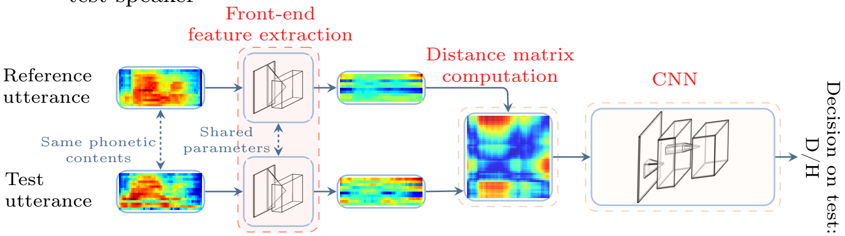
- Distance matrices from optimal representations→ incorporating a front-end feature extraction layer prior to computing distance matrices

- The front-end feature extraction, distance matrix computation, and final healthy and dysarthric speech detection layers are jointly optimized in an end-to-end framework.



CNN-based system exploiting pairwise distance matrices

- System inputs: phonetically-matched pairs of test and reference (healthy) representations
  - Being resized to the same temporal dimension (i.e., by down-sampling and padding)
- Evaluating an utterance from an unseen test speaker
  - Pairing it to its phonetically-matched counterpart from many reference speakers in the training
  - Analysing the given pairs (giving distance matrices) by the CNN classifier
  - Soft voting on all CNN prediction scores for all distance matrices from the test speaker



Outline

Experimental results

» Dataset

- Spanish PC-GITA database (Orozco et al., 2014)
  - 50 **PD patients** vs. 50 healthy speakers (10-fold CV paradigm)
- French MoSpeeDi database
  - 20 **Dysarthric (PD and ALS) patients** vs. 20 healthy speakers (5-fold CV paradigm)

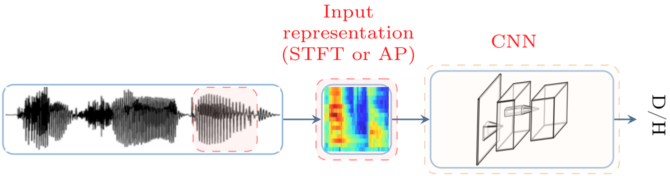
» Evaluation

- Detection accuracy: percentage of correctly classified speakers
- AUC: area under ROC curve

Experimental results

» Baseline networks and the proposed network

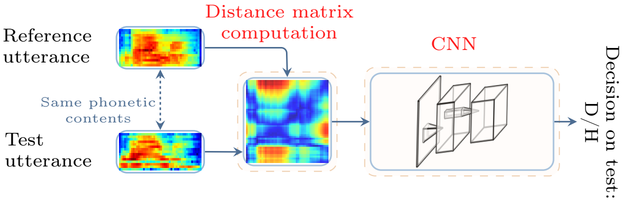
- B-CNN$_1$
- B-CNN$_2$
- Proposed network



Experimental results

» Baseline networks and the proposed network

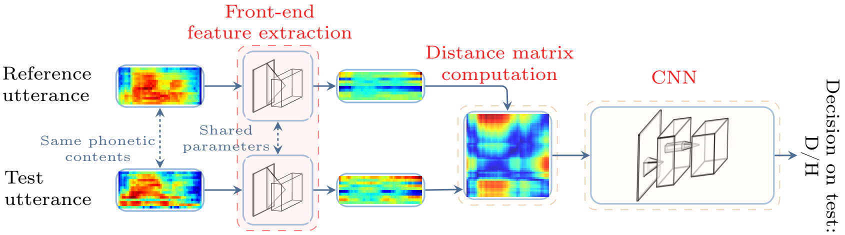
- B-CNN$_1$
- B-CNN$_2$
- Proposed network



Experimental results

» Baseline networks and the proposed network

- B-CNN$_1$
- B-CNN$_2$
- Proposed network



Experimental results

» Classification results using B-CNN_1 on STFT and AP representations for considered databases

Database	Input representation	AUC	Accuracy [%]
Spanish PC-GITA	STFT	0.56	53.67
Spanish PC-GITA	AP	0.75	72.00
French MoSpeeDi	STFT	0.64	52.50
French MoSpeeDi	AP	0.73	60.83

Experimental results

» Classification results using B-CNN$_1$ on STFT and AP representations for considered databases

Database	Input representation	AUC	Accuracy [%]
Spanish PC-GITA	STFT	0.56	53.67
Spanish PC-GITA	AP	0.75	72.00
French MoSpeeDi	STFT	0.64	52.50
French MoSpeeDi	AP	0.73	60.83

- AP representations perform better than STFT independently of the language or diseases
- Demonstrating the advantages of articulatory modeling of speech using AP for dysarthric speech detection

Experimental results

» Classification results using baseline systems and the proposed approach on both databases (input AP representations)

Database	CNN	AUC	Accuracy [%]
Spanish PC-GITA	Baseline B-CNN$_1$	0.75	72.00
Spanish PC-GITA	Baseline B-CNN$_2$	0.78	68.33
Spanish PC-GITA	Proposed	**0.83**	**77.67**
French MoSpeeDi	Baseline B-CNN$_1$	0.733	60.83
French MoSpeeDi	Baseline B-CNN$_2$	0.77	70.83
French MoSpeeDi	Proposed	**0.84**	**76.67**

Experimental results

» Classification results using baseline systems and the proposed approach on both databases (input AP representations)

Database	CNN	AUC	Accuracy [%]
Spanish PC-GITA	Baseline B-CNN_1	0.75	72.00
Spanish PC-GITA	Baseline B-CNN_2	0.78	68.33
Spanish PC-GITA	Proposed	**0.83**	**77.67**
French MoSpeeDi	Baseline B-CNN_1	0.733	60.83
French MoSpeeDi	Baseline B-CNN_2	0.77	70.83
French MoSpeeDi	Proposed	**0.84**	**76.67**

- Proposed approach (pairwise distance-based CNN with a front-end feature extraction layer) significantly improves the performance in comparison to B-CNN_2 (computing distance matrices directly on AP representations)
- Proposed approach outperforms baseline systems for different databases with different languages

Outline

Summary

- Goal: feasibility of automatic dysarthric speech detection using a pairwise distance-based CNN.
- Considering phonetically-matched AP representations from healthy (i.e., reference) and test speakers.
- Extracting features and processing distance matrix computed from features by a CNN-based classifier
- End-to-end optimizing feature extraction, distance matrix computation, and classification.
- The proposed approach is generalizable across languages outperforming state-of-the-art CNN-based systems.

Thank You

Reference

An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., Heitzman, D., and Wang, J. (2018). Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India.

Bhati, S., Velazquez, L. M., Villalba, J., and Dehak, N. (2019). LSTM siamese network for parkinson's disease detection from speech. In *In Proc. IEEE Global Conference on Signal and Information Processing*, pages 1–5, Ottawa, Canada.

Hegde, S., Shetty, S., Rai, S., and Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947.e11–947.e33.

Hernandez, A., Yeo, E. J., Kim, S., and Chung, M. (2020). Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics. In *Proc. 21st Annual Conference of the International Speech Communication Association*, pages 2897–2901, Shanghai, China.

Kodrasi, I. and Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 28(1):1210–1222.

Mallela, J., Illa, A., Belur, Y., Atchayaram, N., Yadav, R., Reddy, P., Gope, D., and Ghosh, P. K. (2020). Raw Speech Waveform Based Classification of Patients with ALS, Parkinson's Disease and Healthy Controls Using CNN-BLSTM. In *Proc. 21st Annual Conference of the International Speech Communication Association*, pages 4586–4590, Shanghai, China.

Orozco, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., and Noeth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Rasipuram, R. and Magimai-Doss, M. (2016). Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech & Language*, 36:233–259.

Vaiciukynas, E., Gelzinis, A., Verikas, A., and Bacauskiene, M. (2017). Parkinson's disease detection from speech using convolutional neural networks. In *In Proc. International Conference on Smart Objects and Technologies for Social Good*, pages 206–215, Pisa, Italy. Springer International Publishing.

Vasquez, J., Orozco, J. R., and Noeth, E. (2017). Convolutional neural network to model articulation impairments in patients with parkinson's disease. In *In Proc. Annual Conference of the International Speech Communication Association*, pages 314–318, Stockholm, Sweden.

Knowledge-based phoneme-to-articulatory feature map

Manner of articulation (degree of constriction), place of articulation (place of constriction), height of articulation (height of the tongue or roundedness) and vowel

Phoneme	Manner	Place	Height	Vowel
sil	sil	sil	sil	sil
aa	vowel	back	low	aa
ae	vowel	mid-front	low	ae
ah	vowel	mid	mid	ah
ao	vowel	back	mid-low	ao
aw1	vowel	mid-front	low	aw1
aw2	vowel	mid-back	high	aw2
ax	vowel	mid	mid	ax
axr	approximant	retroflex	mid	consonant
ay1	vowel	back	low	ay1
ay2	vowel	mid-front	high	ay2
b	voiced-stop	labial	max	consonant
ch	stop	front	max	consonant
d	voiced-stop	alveolar	max	consonant
dh	voiced-fricative	dental	max	consonant
eh	vowel	mid-front	mid	eh
el	approximant	lateral	very-high	consonant
em	nasal	labial	max	consonant
en	nasal	alveolar	max	consonant
er	vowel	mid	mid	er
ey1	vowel	front	mid-high	ey1
ey2	vowel	mid-front	high	ey2
f	fricative	labial	max	consonant
g	voiced-stop	dorsal	max	consonant
hh	aspirated	unknown	max	consonant
ih	vowel	mid-front	high	ih
iy	vowel	front	very-high	iy
jh	voiced-stop	front	max	consonant
k	stop	dorsal	max	consonant
l	approximant	lateral	very-high	consonant
m	nasal	labial	max	consonant
n	nasal	alveolar	max	consonant
ng	nasal	dorsal	max	consonant
ow1	vowel	back	mid	ow1
ow2	vowel	mid-back	high	ow2
oy1	vowel	back	mid-low	oy1
oy2	vowel	mid-front	high	oy2
p	stop	labial	max	consonant
r	approximant	retroflex	mid-low	consonant
s	fricative	alveolar	max	consonant
sh	fricative	front	max	consonant
t	stop	alveolar	max	consonant
th	fricative	dental	max	consonant
uh	vowel	mid-back	high	uh
uw	vowel	back	very-high	uw
v	voiced-fricative	labial	max	consonant
w	approximant	back	very-high	consonant
y	approximant	front	very-high	consonant
z	voiced-fricative	alveolar	max	consonant
zh	voiced-fricative	front	max	consonant