On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches

Guilherme Schu[⋆,†], Parvaneh Janbakhshi[‡], Ina Kodrasi[⋆]

[⋆]Idiap Research Institute, Martigny, Switzerland
[†]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[‡] Bayer AG, Berlin, Germany

ICASSP 2023

June 2023

Outline

Automatic dysarthric speech classification

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS)
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness

Automatic dysarthric speech classification

- Dysarthria of speech $\rightarrow$ disturbances of muscular control on speech production system
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS)
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness
- Dysarthric speech classification: discriminating between speech from healthy and dysarthric speakers

Automatic dysarthric speech classification

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS)
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness
- Dysarthric speech classification: discriminating between speech from healthy and dysarthric speakers

Dysarthric speech classification using:

- Subjective screening based on judgment of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis
- Automatic and objective method
  - Efficient and economical
  - Repeatable
  - Early diagnosis

Automatic dysarthric speech classification

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS)
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness
- Dysarthric speech classification: discriminating between speech from healthy and dysarthric speakers

Dysarthric speech classification using:

- Subjective screening based on judgment of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis
- Automatic and objective method
  - Efficient and economical
  - Repeatable
  - Early diagnosis

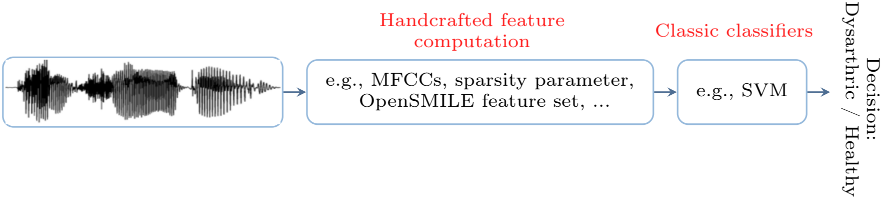
Outline

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

- Traditional machine learning approaches
- Deep learning approaches

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

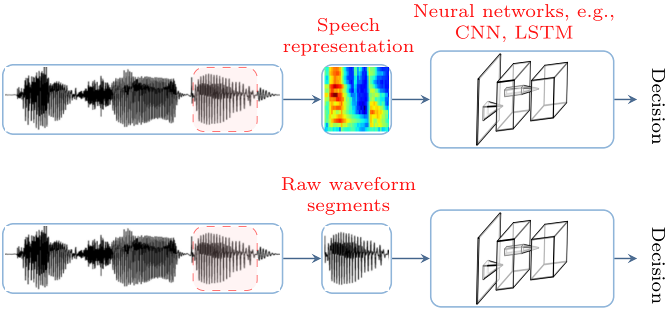
- Traditional machine learning approaches (Hegde et al., 2019; Kodrasi and Bourlard, 2020; Hernandez et al., 2020; Narendra and Alku, 2018)



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

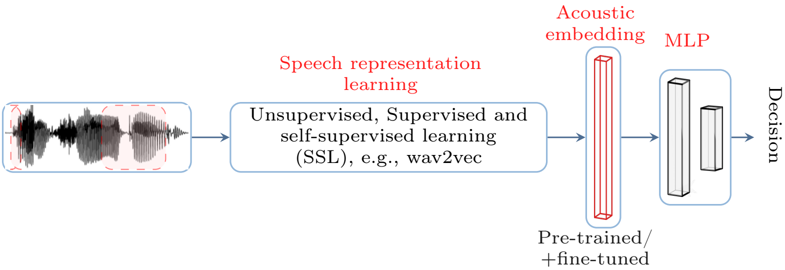
» Deep learning approaches → data-driven approaches using no prior knowledge

- Exploit high-level abstract features from low-level speech representations or raw waveforms using neural networks (Vaiciukynas et al., 2017; Mallela et al., 2020; Narendra et al., 2021)



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

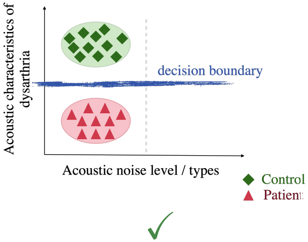
- Deep learning approaches → data-driven approaches using no prior knowledge
  - Speech representation (embedding) learning + downstream task, i.e., dysarthric speech classification (Janbakhshi and Kodrasi, 2021; Yang et al., 2021; Janbakhshi and Kodrasi, 2022)

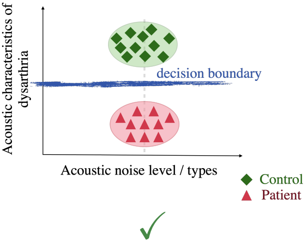


Underlying assumption

» Assumptions in state-of-the-art automatic dysarthria classification systems

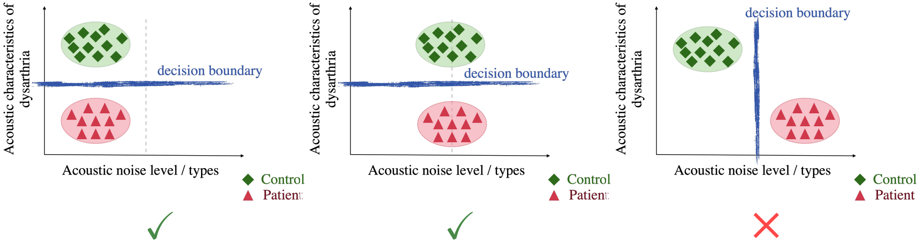
- Control and dysarthric speakers are recorded in the same (ideally noiseless) environment using the same recording setup. ✓





Underlying assumption

- Assumptions in state-of-the-art automatic dysarthria classification systems
  - Control and dysarthric speakers are recorded in the same (ideally noiseless) environment using the same recording setup. ✓
- Consistent violation of the assumptions across the speaker groups
  - Trained classifiers would learn characteristics of the recording environment instead of dysarthric speech characteristic. ✗



Underlying assumption

- Assumptions in state-of-the-art automatic dysarthria classification systems
  - Control and dysarthric speakers are recorded in the same (ideally noiseless) environment using the same recording setup. ✓
- Consistent violation of the assumptions across the speaker groups
  - Trained classifiers would learn characteristics of the recording environment instead of dysarthric speech characteristic. ✗
- Some of publicly available datasets, validated in many state-of-the-art automatic dysarthria classification approaches, might not fulfill such assumptions!
  - UA-Speech (Rudzicz et al., 2012) and TORGO (Kim et al., 2008) ✗

Outline

Investigating the recording environment bias on dysarthric speech classification

Proposed

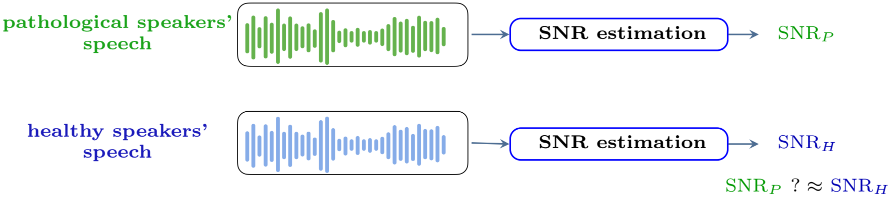
Investigating if the dysarthria classification results using the UA-Speech and TORGO databases reflect the characteristics of the recording environment rather than characteristics of dysarthric speech.

Investigating the recording environment bias on dysarthric speech classification

Proposed

Investigating if the dysarthria classification results using the UA-Speech and TORGO databases reflect the characteristics of the recording environment rather than characteristics of dysarthric speech.

(1) assessing variability in recording conditions using signal-to-noise ratio (SNR) estimation

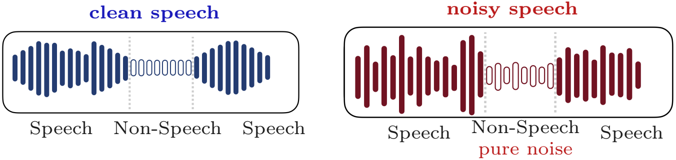


Investigating the recording environment bias on dysarthric speech classification

Proposed

Investigating if the dysarthria classification results using the UA-Speech and TORGO databases reflect the characteristics of the recording environment rather than characteristics of dysarthric speech.

(2) assessing state-of-the-art dysarthria classification approaches on speech-only and non-speech-only segments

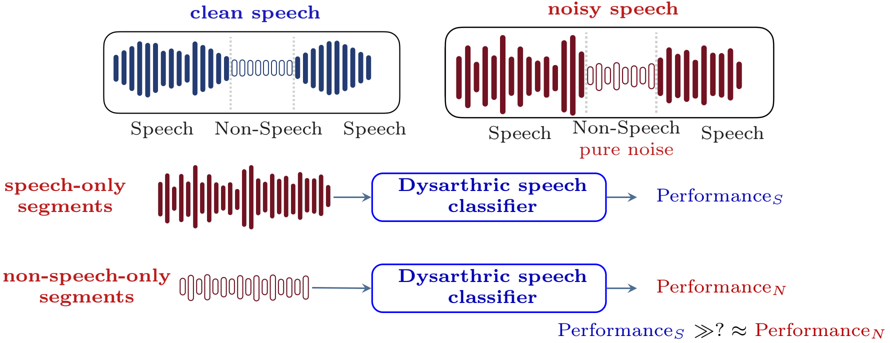


Investigating the recording environment bias on dysarthric speech classification

Proposed

Investigating if the dysarthria classification results using the UA-Speech and TORGO databases reflect the characteristics of the recording environment rather than characteristics of dysarthric speech.

(2) assessing state-of-the-art dysarthria classification approaches on speech-only and non-speech-only segments



Outline

Evaluation

- Dataset 1: English UA-Speech (Rudzicz et al., 2012)
  - Discriminating 15 dysarthric (CP) patients from 13 healthy speakers
  - Considering recordings of 721 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework
- Dataset 2: English TORGO (Kim et al., 2008)
  - Discriminating 7 dysarthric (CP or ALS) patients from 7 healthy speakers
  - Considering recordings of 62 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework

Evaluation

» Dataset 1: English UA-Speech (Rudzicz et al., 2012)
  - Discriminating 15 dysarthric (CP) patients from 13 healthy speakers
  - Considering recordings of 721 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework

» Dataset 2: English TORGO (Kim et al., 2008)
  - Discriminating 7 dysarthric (CP or ALS) patients from 7 healthy speakers
  - Considering recordings of 62 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework

» Applied VAD: forced alignment from a ASR system (Hermann and Magimai-Doss, 2021) to extract speech and non-speech segments

Evaluation

- Dataset 1: English UA-Speech (Rudzicz et al., 2012)
  - Discriminating 15 dysarthric (CP) patients from 13 healthy speakers
  - Considering recordings of 721 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework
- Dataset 2: English TORGO (Kim et al., 2008)
  - Discriminating 7 dysarthric (CP or ALS) patients from 7 healthy speakers
  - Considering recordings of 62 phonetically-matched utterances per speaker
  - Leave-one-speaker-out validation framework
- Applied VAD: forced alignment from a ASR system (Hermann and Magimai-Doss, 2021) to extract speech and non-speech segments
- Evaluation metric for dysarthric speech classification
  - Speaker-level classification accuracy

SNR of control and dysarthric recordings

- SNR estimation
  - Using a data-driven recurrent neural network to estimate utterance-level SNR (Li et al., 2021)

Mean and standard deviation of the estimated SNRs [dB] across all utterances of control and dysarthric speakers in the UA-Speech and TORGO databases.

Speakers	UA-Speech	TORGO
Control	3.7 ± 11.5	2.1 ± 13.2
Dysarthric	-7.6 ± 16.1	-4.0 ± 14.7

SNR of control and dysarthric recordings

» SNR estimation

- Using a data-driven recurrent neural network to estimate utterance-level SNR (Li et al., 2021)

Mean and standard deviation of the estimated SNRs [dB] across all utterances of control and dysarthric speakers in the UA-Speech and TORGO databases.

Speakers	UA-Speech	TORGO
Control	3.7 ± 11.5	2.1 ± 13.2
Dysarthric	-7.6 ± 16.1	-4.0 ± 14.7

- Large variation in the acoustic conditions of the recorded utterances for both databases
- Large difference in the average SNRs of control and dysarthric utterances in both databases
- No guarantee that automatic dysarthria classification approaches validated on these databases learn speech characteristics or recording conditions changes for the two groups of speakers.

Dysarthria classification approaches

- SVM classifier trained on hand crafted features
  - OpenSMILE feature set + PCA dimensionality reduction
  - MFCCs functionals (48-dimensional feature vectors)
  - Sparsity-based features (129-dimensional feature vectors)

Dysarthria classification approaches

- SVM classifier trained on hand crafted features
  - OpenSMILE feature set + PCA dimensionality reduction
  - MFCCs functionals (48-dimensional feature vectors)
  - Sparsity-based features (129-dimensional feature vectors)
- Convolutional neural networks (CNNs)
  - Operating on Mel-scale input spectrograms (Vásquez-Correa et al., 2017)

Dysarthria classification approaches

- SVM classifier trained on hand crafted features
  - OpenSMILE feature set + PCA dimensionality reduction
  - MFCCs functionals (48-dimensional feature vectors)
  - Sparsity-based features (129-dimensional feature vectors)
- Convolutional neural networks (CNNs)
  - Operating on Mel-scale input spectrograms (Vásquez-Correa et al., 2017)
- Speech representation learning (SRL)
  - A supervised auto-encoder operating on Mel-scale input spectrograms (Janbakhshi and Kodrasi, 2021)

Dysarthria classification approaches

- SVM classifier trained on hand crafted features
  - OpenSMILE feature set + PCA dimensionality reduction
  - MFCCs functionals (48-dimensional feature vectors)
  - Sparsity-based features (129-dimensional feature vectors)
- Convolutional neural networks (CNNs)
  - Operating on Mel-scale input spectrograms (Vásquez-Correa et al., 2017)
- Speech representation learning (SRL)
  - A supervised auto-encoder operating on Mel-scale input spectrograms (Janbakhshi and Kodrasi, 2021)
- Wav2vec2 learned representation + MLP classifier
  - MLP trained on wav2vec2 embeddings with or without fine-tuning (Yang et al., 2021)

Dysarthria classification [mean and standard deviation of the accuracy]

Approach validated on **UA-Speech**	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	81.0 ± 19.8	84.5 ± 21.9	83.3 ± 21.1
SVM+MFCCs	81.0 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
SVM+sparsity-based features	94.0 ± 1.7	96.4 ± 0.0	96.4 ± 0.0
CNN+Mel spectrograms	95.2 ± 1.7	97.6 ± 1.7	98.8 ± 1.7
SRL+Mel spectrograms	98.8 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
MLP+ft-wav2vec2	95.2 ± 1.7	97.6 ± 1.7	95.2 ± 1.7
MLP+wav2vec2	54.8 ± 1.7	58.3 ± 1.7	54.8 ± 1.7

Approach validated on **TORGO**	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	60.0 ± 5.4	82.2 ± 6.3	71.1 ± 12.6
SVM+MFCCs	60.0 ± 0.0	88.9 ± 3.1	57.8 ± 3.1
SVM+sparsity-based features	73.3 ± 0.0	93.3 ± 0.0	73.3 ± 5.4
CNN+Mel spectrograms	53.3 ± 11.5	77.8 ± 10.2	68.9 ± 10.2
SRL+Mel spectrograms	71.1 ± 3.1	100.0 ± 0.0	91.1 ± 3.1
MLP+ft-wav2vec2	60.0 ± 5.4	57.8 ± 3.1	60.0 ± 5.4
MLP+wav2vec2	55.6 ± 3.1	57.8 ± 3.1	57.8 ± 6.3

Dysarthria classification [mean and standard deviation of the accuracy]

Approach validated on **UA-Speech**	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	81.0 ± 19.8	84.5 ± 21.9	83.3 ± 21.1
SVM+MFCCs	81.0 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
SVM+sparsity-based features	94.0 ± 1.7	96.4 ± 0.0	96.4 ± 0.0
CNN+Mel spectrograms	95.2 ± 1.7	97.6 ± 1.7	98.8 ± 1.7
SRL+Mel spectrograms	98.8 ± 1.7	100.0 ± 0.0	100.0 ± 0.0
MLP+ft-wav2vec2	95.2 ± 1.7	97.6 ± 1.7	95.2 ± 1.7
MLP+wav2vec2	54.8 ± 1.7	58.3 ± 1.7	54.8 ± 1.7

Approach validated on **TORGO**	Speech	Non-speech	Speech&Non-speech
SVM+openSMILE	60.0 ± 5.4	82.2 ± 6.3	71.1 ± 12.6
SVM+MFCCs	60.0 ± 0.0	88.9 ± 3.1	57.8 ± 3.1
SVM+sparsity-based features	73.3 ± 0.0	93.3 ± 0.0	73.3 ± 5.4
CNN+Mel spectrograms	53.3 ± 11.5	77.8 ± 10.2	68.9 ± 10.2
SRL+Mel spectrograms	71.1 ± 3.1	100.0 ± 0.0	91.1 ± 3.1
MLP+ft-wav2vec2	60.0 ± 5.4	57.8 ± 3.1	60.0 ± 5.4
MLP+wav2vec2	55.6 ± 3.1	57.8 ± 3.1	57.8 ± 6.3

- Performance of majority of approaches using non-speech segments is the same or even better than when using speech segments or complete utterances from the UA-Speech and TORGO databases
- Classification results obtained on the UA-Speech and TORGO databases can be greatly affected by characteristics of the recording environment and setup

Outline

Summary

- Investigating the use of the UA-Speech and TORGO databases to validate automatic dysarthria classification approaches

- Hypothesizing that classification results could be biased towards capturing characteristics of the recording environment rather than characteristics of dysarthric speech
  - Estimating the utterance-level SNRs on these databases
  - Validating state-of-the-art dysarthria classification approaches on the speech and non-speech segments of these database

- Experimental results have shown that:
  - Utterance-level SNRs in control and dysarthric recordings are considerably different in both databases
  - State-of-the-art approaches achieve the same or a better dysarthria classification performance when using only the non-speech segments than when using only the speech segments.

- Awareness on the bias of recordings quality in validating classification approaches

Thank You

Reference I

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.

Hermann, E. and Magimai-Doss, M. (2021). Handling acoustic variation in dysarthric speech recognition systems through model combination. In *Proc. Annual Conference of the International Speech Communication Association*, pages 4788–4792, Brno, Czechia.

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.

Janbakhshi, P. and Kodrasi, I. (2021). Supervised speech representation learning for Parkinson's disease classification. In *Proc. ITG conference on Speech Communication*, pages 154–158, Kiel, Germany.

Janbakhshi, P. and Kodrasi, I. (2022). Adversarial-Free Speaker Identity-Invariant Representation Learning for Automatic Dysarthric Speech Classification. In *Proc. Interspeech 2022*, pages 2138–2142.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Proc. Annual Conference of the International Speech Communication Association*, pages 1741–1744, Brisbane, Australia.

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.

Li, H., Wang, D., Zhang, X., and Gao, G. (2021). Recurrent neural networks and acoustic features for frame-level signal-to-noise ratio estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2878–2887.

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.

Narendra, N., Schuller, B., and Alku, P. (2021). The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936.

Narendra, N. P. and Alku, P. (2018). Dysarthric speech classification using glottal features computed from non-words, words and sentences. In *Proc. Annual Conference of the International Speech Communication Association*, pages 3403–3407, Hyderabad, India.

Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541.

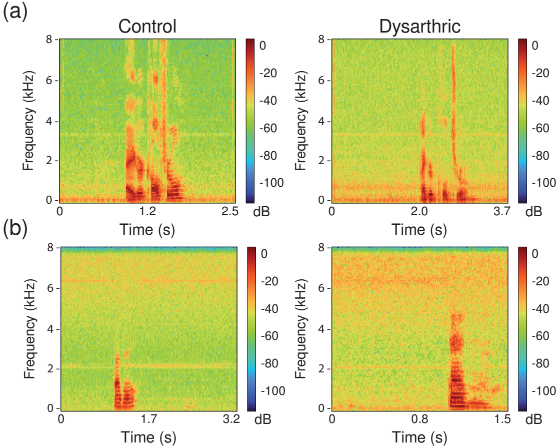
Reference II

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.

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

Yang, S., Chi, P., Chuang, Y., Lai, C. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G., et al. (2021). Superb: Speech processing universal performance benchmark. In *Proc. Annual Conference of the International Speech Communication Association*, pages 1194–1198, Brno, Czechia.

UA-Speech and TORGO databases



Spectrograms of an exemplary utterance from a control and dysarthric speaker from the a) UA-Speech and b) TORGO databases.