# Phonet: a Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech

J. C. Vásquez-Correa[1,2], P. Klumpp[1], J. R. Orozco-Arroyave[2] and E. Nöth[1]

[1]*Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*

[2]*GITA research Lab, Faculty of Engineering, University of Antioquia, UdeA, Medellín, Colombia*

INTERSPEECH 2019

## Introduction

- High-dimensional feature sets like MFCCs or embeddings from neural networks are rarely used for medical applications due to their lack of interpretability.
- Phonological features can be more comprehensible for clinicians than the traditional acoustic features used in speech processing.
- Phonological features are represented by a vector with explainable information about the mode and manner of articulation of the speaker.
- These features are commonly understood by clinicians since they are related with movements of the articulators in the vocal tract.

**Aim:**

1. To map high-dimensional feature vectors into explainable feature vectors called **phonological posteriors** that can be comprehensible for the medical community.
2. To create an open source toolkit to estimate phonological posteriors based on bidirectional RNNs with gated recurrent units (GRUs).

## Phonet

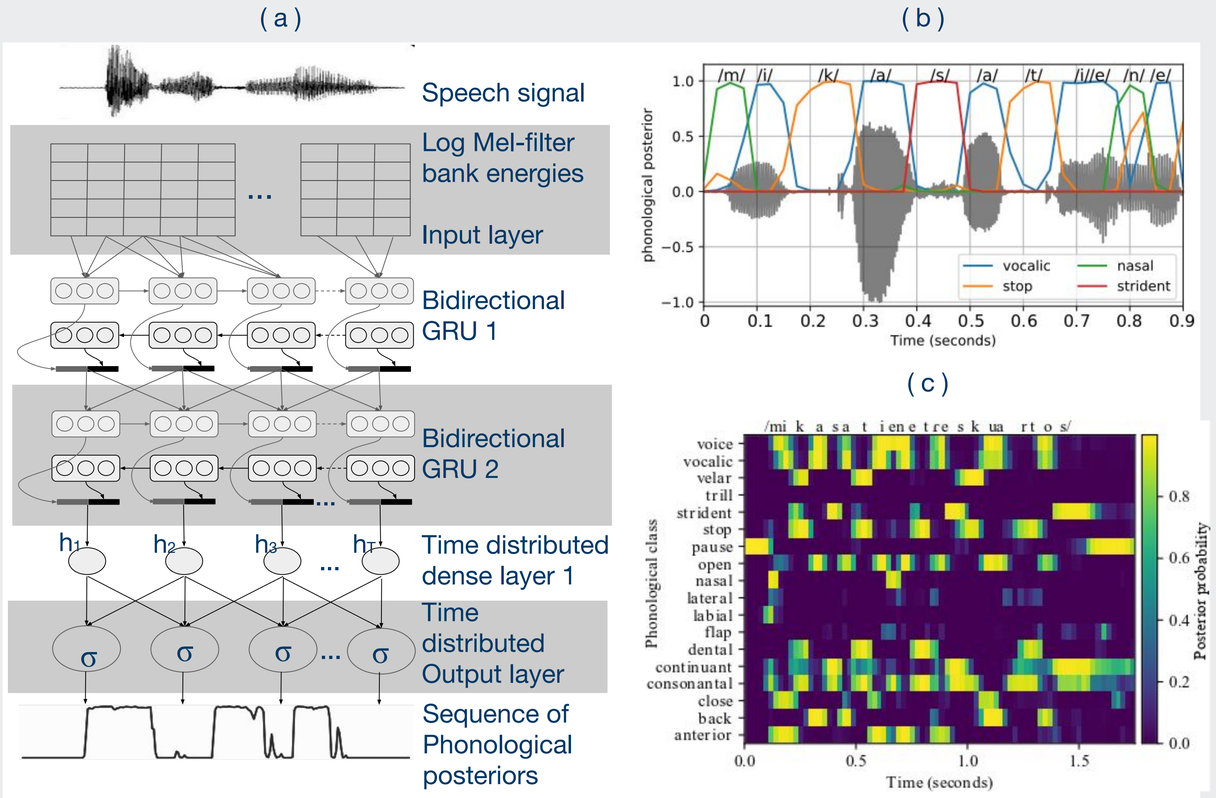


**Figure 1: (a)** Architecture of Phonet, **(b)** Example of the *vocalic, stop, nasal, and strident* phonological posteriors estimated for the sentence “mi casa tiene” in Spanish language. **(c)** Posteriorgram obtained for the Spanish sentence “mi casa tiene tres cuartos”.

## Phonological posteriors

- The phonetic units of a language can be grouped into phonological classes based on the mode and manner of articulation of the sounds.
- 18 Phonological classes considered in this study.

| Phonological classes | | |
|---|---|---|
| 1. Vocalic | 7. Nasal | 13. Lateral |
| 2. Consonantal | 8. Stop | 14. Strident |
| 3. Back | 9. Continuant | 15. Labial |
| 4. Anterior | 10. Flap | 16. Dental |
| 5. Open | 11. Trill | 17. Velar |
| 6. Close | 12. Voiced | 18. Pause |

**Table 1:** List of phonological classes.

- The phonological posteriors will be the posterior probability of a speech frame to belong to one or more phonological classes.
- The phonological posteriors are estimated with a bank of parallel RNNs.
- An additional model for phoneme recognition is also included.

## Data

- CIEMPIESS corpus [1].
- 17 hours of FM podcasts in Mexican Spanish.
- The corpus was forced-aligned at phoneme level [2].
- The aligned phonemes were used as labels to train models for phoneme recognition and for the estimation of the phonological posteriors.

## Results

- Estimation of Phonological posteriors

| Class | F-score | Class | F-score | Class | F-score |
|---|---|---|---|---|---|
| 1. Vocalic | 0.841 | 7. Nasal | 0.907 | 13. Lateral | 0.915 |
| 2. Cons. | 0.831 | 8. Stop | 0.855 | 14. Strident | 0.932 |
| 3. Back | 0.882 | 9. Continuant | 0.861 | 15. Labial | 0.885 |
| 4. Anterior | 0.889 | 10. Flap | 0.895 | 16. Dental | 0.898 |
| 5. Open | 0.841 | 11. Trill | 0.986 | 17. Velar | 0.930 |
| 6. Close | 0.901 | 12. Voiced | 0.885 | 18. Pause | 0.958 |

**Table 2:** Recognition of the phonological classes using Phonet

- Phoneme recognition

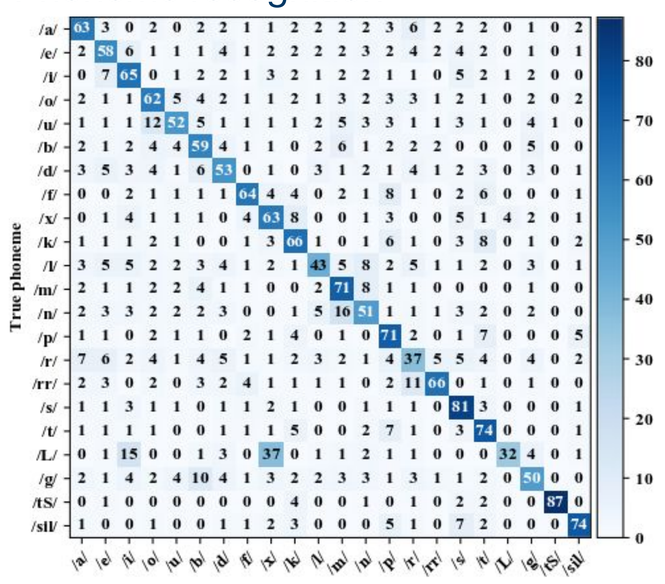


**Figure 2:** Normalized confusion matrix (in %) for the phoneme recognition task using the proposed approach. The blue bar indicates the percentage of samples classified for each phoneme.

## Conclusion

- The accuracy of the models ranges from 80.4% to 93.3%, depending on the phonological class.
- Phonet was trained to recognize posteriors in Spanish. However, the training process can be adapted to other languages.
- Future models will include the estimation of phonological posteriors for the English and German languages.
- The trained models are available as an open-source toolkit.

## References

**[1]** Hernández-Mena, C. D., et. al. (2014). CIEMPIESS: A new open-sourced Mexican Spanish radio corpus. In *LREC* (pp. 371-375).

**[2]** Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *ICPhS* (pp. 371-375).

https://github.com/jcvasquezc/phonet

## Contact

**Camilo Vasquez**
Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

juan.vasquez@fau.de
+49 9131 85 27137
@jcvasquezc1

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287.