Principal Components Analysis of amplitude envelopes from spectral channels: comparison between music and speech.

Duniec, Agnieszka Crouzet, Olivier Delais-Roussarie, Elisabeth

Laboratoire de Linguistique de Nantes – LLING / UMR6310, Nantes Université / CNRS, France

Keep informed / Download

P36

Introduction

- According to the efficient coding hypothesis [1], sensory systems have evolved to encode environmental signals in an optimal way following information theory;
- This would provide a way to represent the greatest amount of information at the lowest possible cost in terms of resources;
- Possible implications for optimal cochlear implant boundary determination [2].

Previous work on speech: Ming and Holt (2009)

- Identification of 6-channel vocoded speech is overall better with "efficient-coding" based frequency boundaries than with logarithmically ordered cochleotopic boundaries;
  - Overall superiority for word recognition in sentences and phoneme identification in non-words;

Previous work on speech: Ueda and Nakajima (2017)

- Factor Analysis of speech signals in 8 different languages [4];
  - 20 frequency channels;
  - Determining the number of "optimal" channels: 4 "optimal" channels;
  - Locating the boundaries between these "optimal" channels: the boundaries would be stable whatever the language studied;
  - Authors do not provide grounded arguments in favor of the number of 4 channels (in contrast to 3, 5, 6, 7. . .).

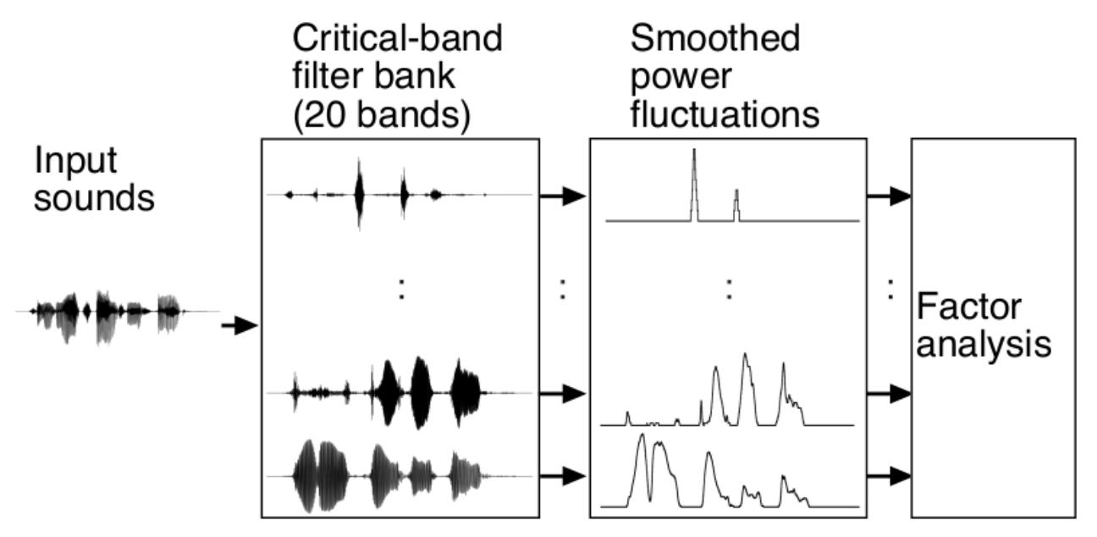
Input sounds
Critical-band filter bank (20 bands)
Smoothed power fluctuations
Factor analysis

Figure 1: Schematic data processing diagram [4]

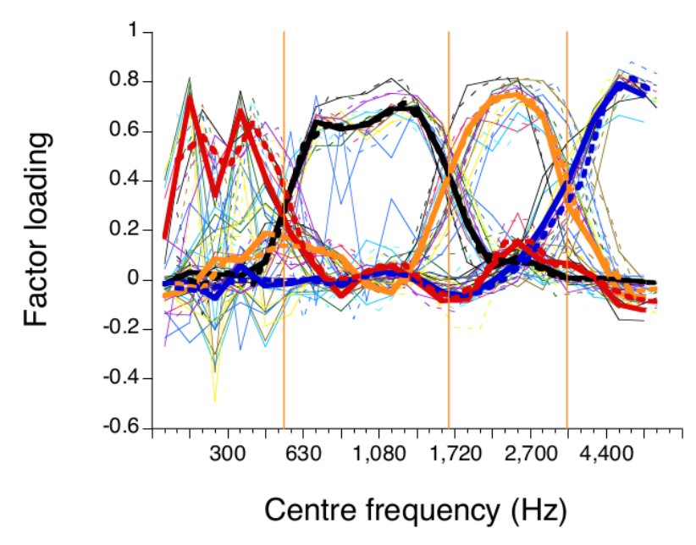

Figure 2: Factor loading curves for speech with 4 Principal Components (Ueda and Nakajima [4]).

Grange and Culling (2018)

- Around 100 frequency channels;
  - English language recordings only;
  - Similar process for the extraction of natural sound statistics;
  - Added comparison with perceptual results on simulations of cochlear implants.

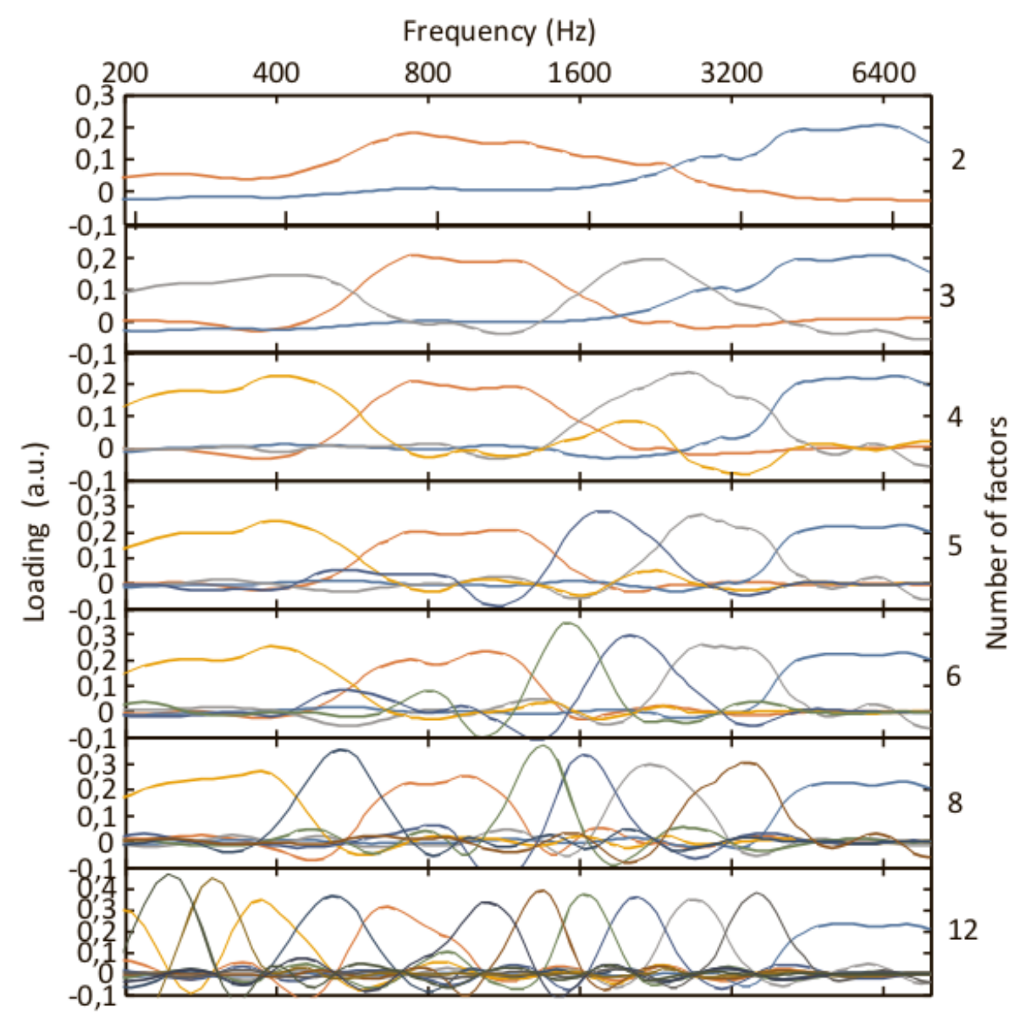

Figure 3: Factor loading curves depending on the number of PCs for speech (Grange and Culling [2]).

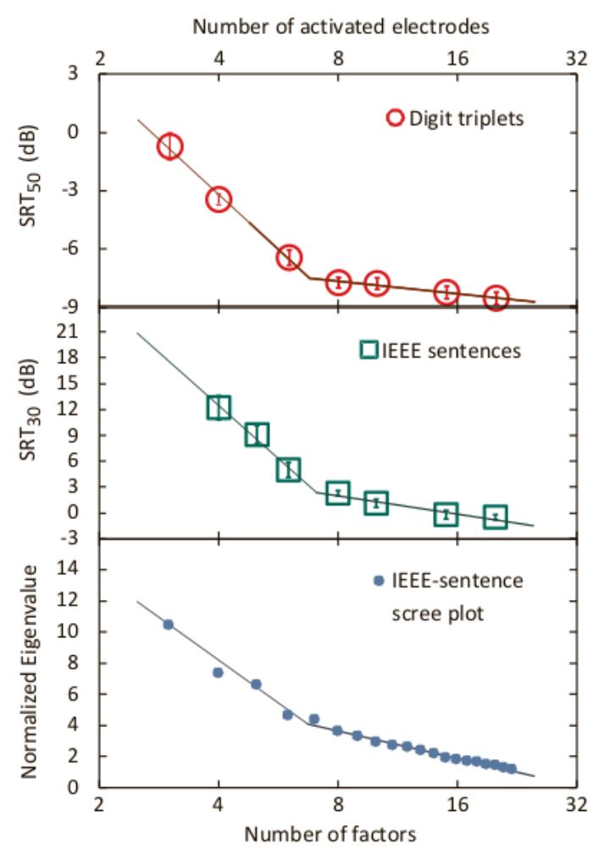

Figure 4: Scree plot (bottom): percentage of explained variance associated with each PC. The inflection point in the scree plot is compared with the perceptual performance data in terms of identification threshold for both digit triplets (top) and simple sentences (middle) (Grange and Culling [2]).

References

[1] E. C. Smith and M. S. Lewicki. "Efficient auditory coding". In: *Nature* 439.7079 (2006), pp. 978–982. — [2] J. Grange and J. Culling. "The Factor Analysis of Speech: Limitations and Opportunities for Cochlear Implants". In: *Acta Acustica united with Acustica* 104 (Sept. 2018), pp. 835–838. — [3] V. L. Ming and L. L. Holt. "Efficient coding in human auditory perception". In: *The Journal of the Acoustical Society of America* 126.3 (Sept. 2009), pp. 1312–1320. — [4] K. Ueda and Y. Nakajima. "An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech". In: *Scientific Reports* 7 (Feb. 2017), p. 42468. — [5] J. J. Galvin, Q.-J. Fu, and R. V. Shannon. "Melodic Contour Identification and Music Perception by Cochlear Implant Users". In: *Annals of the New York Academy of Sciences* 1169.1 (2009), pp. 518–533. — [6] J. D. Crew, J. J. Galvin, and Q.-J. Fu. "Melodic contour identification and sentence recognition using sung speech". In: *The Journal of the Acoustical Society of America* 138.3 (2015), EL347–EL351. — [7] M. Defferrard et al. "FMA: Dataset For Music Analysis". In: *18th International Society for Music Information Retrieval Conference* (Sept. 2017). — [8] S. Graetzer et al. "Dataset of British English speech recordings for psychoacoustics and speech processing research: The Clarity Speech Corpus". In: *Data in Brief* (). —

Acknowledgements

Agnieszka Duniec receives PhD funding (2019–2023) from the RFI-Ouest Industries Créatives (RFI-OIC, Région Pays de la Loire) & Nantes Université.

Aims

1. Research focus 1:
   - Performance observed on vocoded signal material in normal-hearing listeners as well as in CI users is systematically better for speech signals than for music [5, 6];
   - Our aim is to compare statistical properties of music and speech signals in order to evaluate their respective contributions to the efficient coding theoretical proposal.
2. Research focus 2:
   - The idea of fixed boundaries that would not depend on the type of natural signals (e.g. speech vs. music, differences between speakers / between instruments) seems unlikely;
   - Hence the need to be able to assess this variation;
   - Our aim is to propose an objective method for determining the spectral boundaries as a function of the acoustic signal.

Music samples

- Music samples from the FMA open source database [7] (Free Music Archive, https://github.com/mdeff/fma, MP3 files);
- The corpus size for analyses is 400 random samples (10 s. duration each) out of the 8,000 recordings available in the smallest version of the FMA database (approx. 4000 s);
- The mp3 compression level varies between 128 and 256 kbits/s.

Speech samples

- A free corpus of speech signals [8] (Clarity Speech, https://doi.org/10.17866/rd.salford.16918180);
- A random sample of 1,600 out of 10,000 sentences (approx. 4,500 s) from the British National Corpus (BNC), produced by 40 speakers of British English ;
- All audio files are stored in single channel 32-bit floating point wav format at a 44.1kHz sampling rate;

Procedure

1. Principal Components Analysis.
   - For both, music and speech, signal processing and statistical procedures were carried out in the Matlab environment and were mirrored from previous studies on speech [2, 4];
   - As our aim was to compare speech and music, for which typical signal bandwidths differ, two higher-frequency limits were compared (8000 Hz vs. 16000 Hz).
2. Automated estimation of boundary location
   - **Identification of the adjacent curves** by relating the frequency of the peak and the rank of the Principal component ;
   - **Estimates of the intersection**: matching of adjacent curves in the spectrum and averaging based on (1) the lower boundary estimate for the upper channel and (2) the upper boundary estimate for the lower channel (Fig. 5);

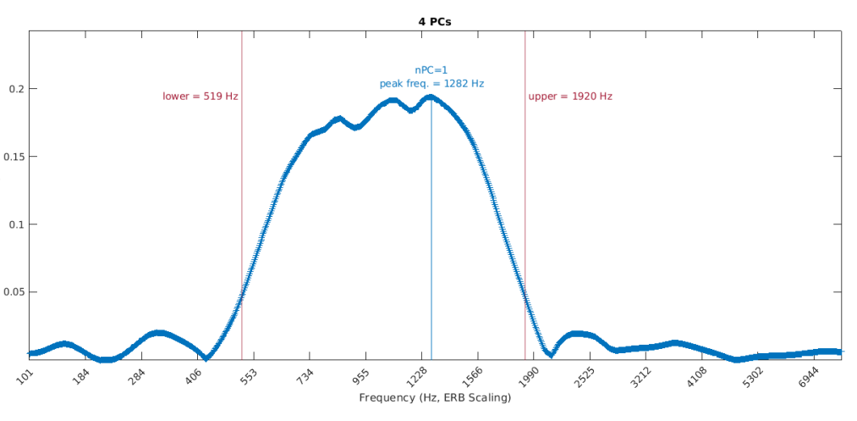

Figure 5: Initial broad boundary estimation.

   - **Modelling the intersection** using first-order polynomials (straight lines, Fig. 6).

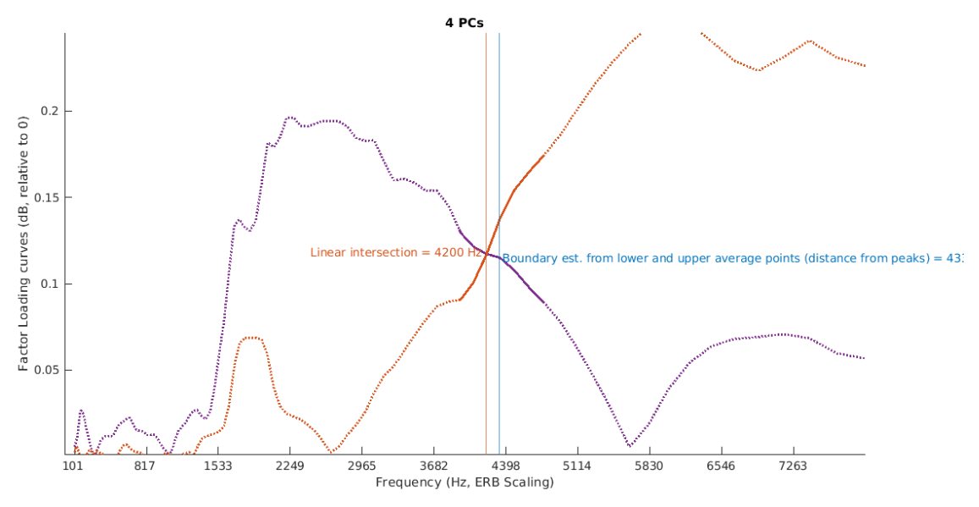

Figure 6: Final boundary estimation from intersection modelling.

	Ch1/Ch2	Ch2/Ch3	Ch3/Ch4
Our initial estimate (in Hz)	586	1752	3775
Our estimate using linear modelling (in Hz)	587	1735	3745
Ueda & Nakajima estimate (2017, en Hz)	540	1720	3300

	Ch1/Ch2	Ch2/Ch3	Ch3/Ch4
Difference observed (in semitones)	1.44	0.15	2.19
Difference observed (in Barks)	0.39	0.06	0.78

Results and Discussion

1. Speech signals
   - In accordance with Grange and Culling [2], there's an inflection point in the scree plot for speech signals;
   - Our estimates of frequency boundaries identified from speech do not closely match those of Ueda and Nakajima [4];
2. Music signals
   - Contrary to Grange and Culling [2]: no inflection point in the scree plot;
3. Music vs. Speech
   - Boundaries are not fixed and depend on the type of natural signals (speech vs. music).
   - Observations on variation in (1) boundary location and (2) PC Rank/frequency relations in figs. 11 to 13. . .
4. **Work in progress:**
   - Replicate findings on a database of single instrument music recordings in order to get conceptually closer to "clean speech".
   - Behavioral study on vocoded signals is currently being prepared in order to evaluate how these measures would impact perceptual processing for speech and melodies.

Factor loadings and Scree plot for a 1 hour music sample

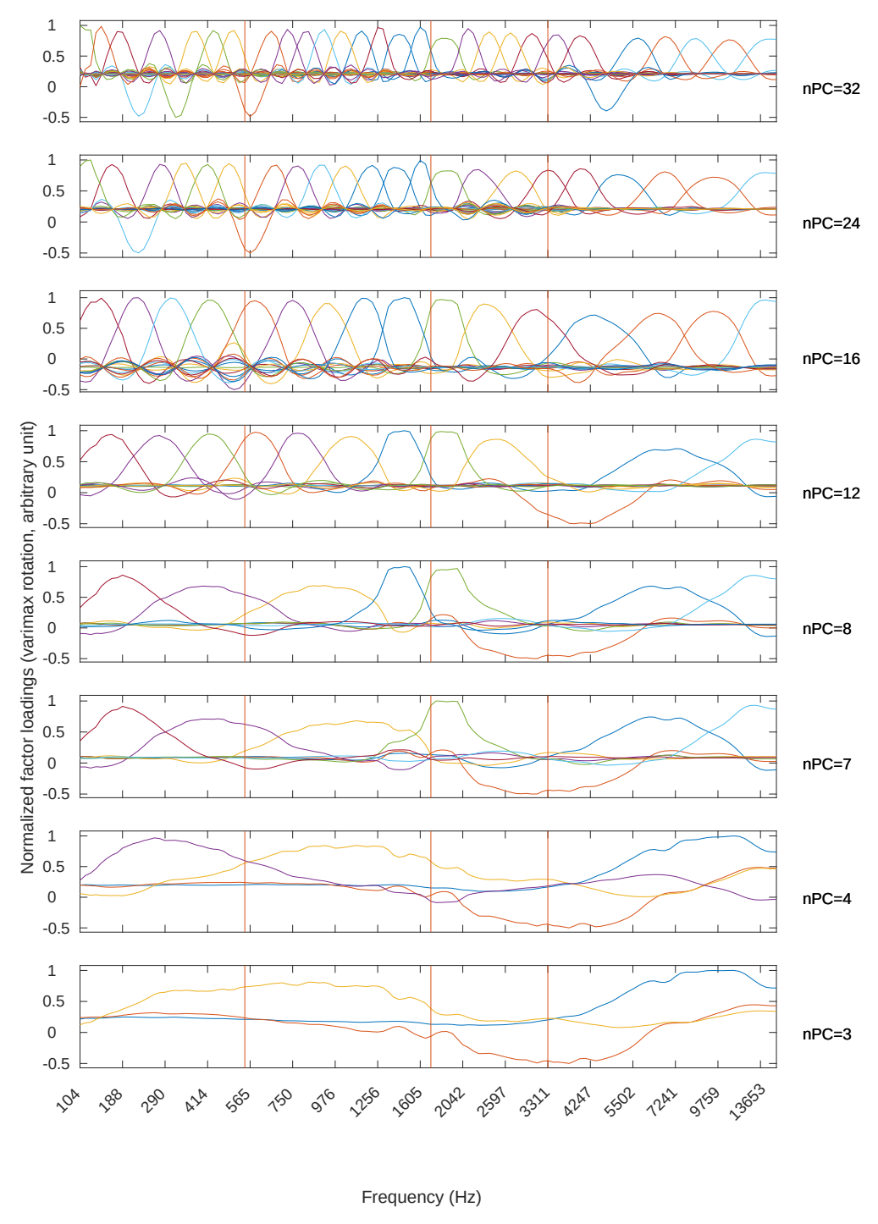

Figure 7: Factor loading curves for 1 hour of random music samples, frequency boundaries identified by Ueda and Nakajima [4] are epresented as orange vertical segments (High-Frequency boundary: 16 kHz).

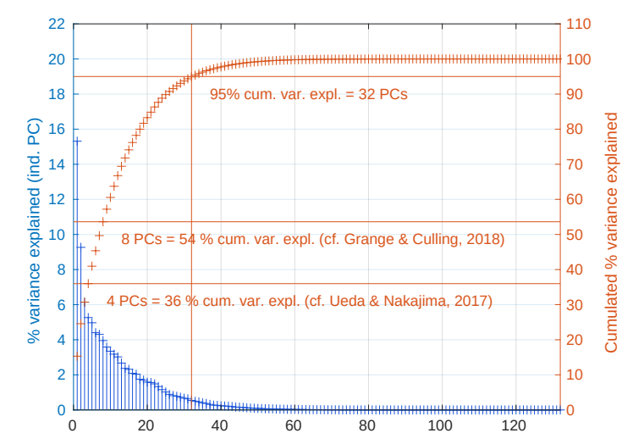

Figure 8: Scree plot (blue: individual values of explained variance, red: cumulated percentage of explained variance) for 1 hour of music with 16 kHz frequency limit. Possible targets: Inflection point (none identified) or a given % (here, the 95% boundary is indicated).

Factor loadings and Scree plot for a 1 hour speech sample

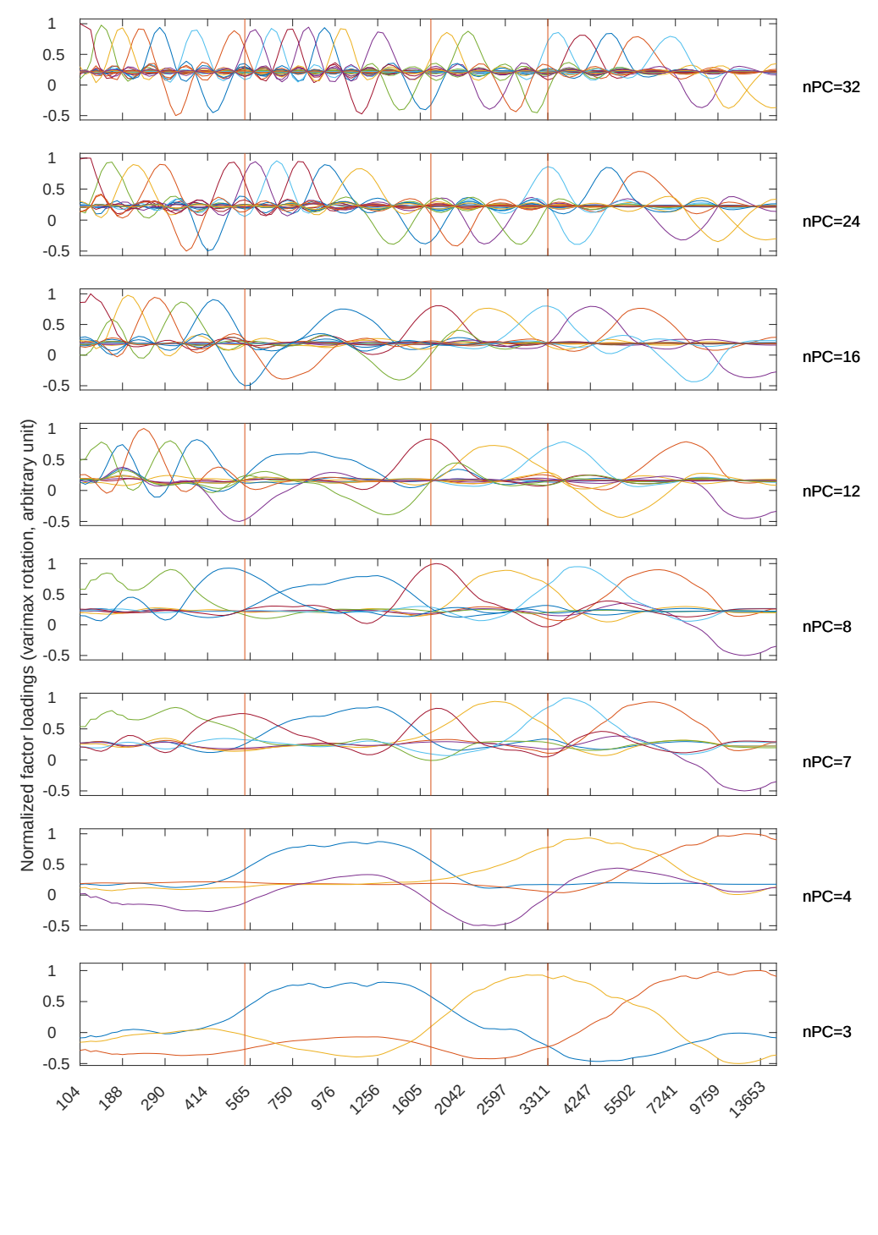

Figure 9: Factor loading curves for 1 hour of random speech samples, frequency boundaries identified by Ueda and Nakajima [4] are represented as orange vertical segments (High-Frequency boundary: 16 kHz).

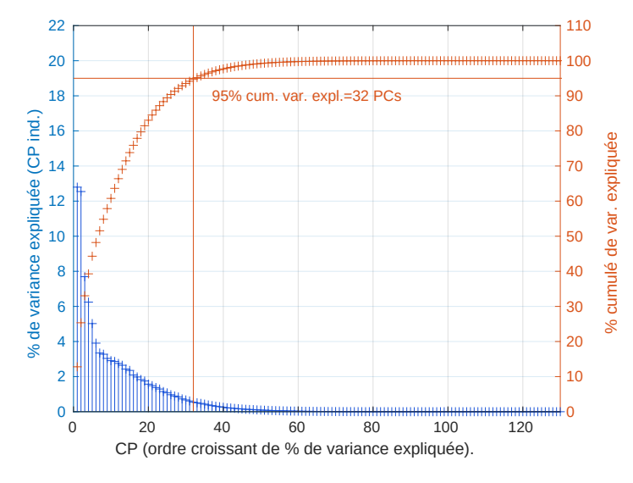

Figure 10: Scree plot (blue: individual values of explained variance, red:cumulated percentage of explained variance) for 1 hour of speech with 16 kHz frequency limit. Possible targets: Inflection point (at 7th or 8th PCs) or a given % (here, the 95% boundary is indicated).

Boundary comparison / music vs speech, 8kHz

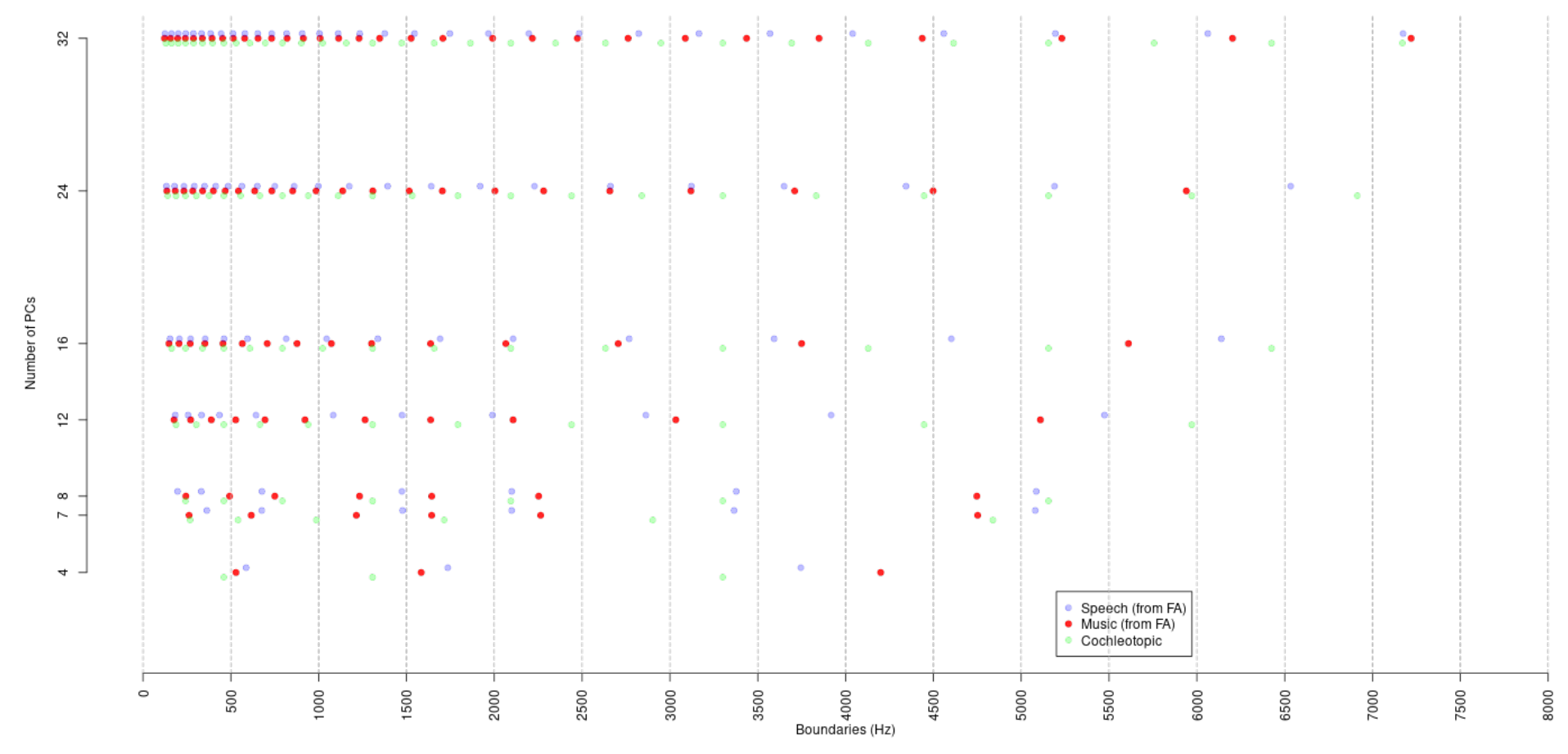

Figure 11: Boundary comparison (blue: speech boundaries from FA, red:music boundaries from FA, green: logarithmic boundaries)

Boundary comparison music vs speech, 16kHz

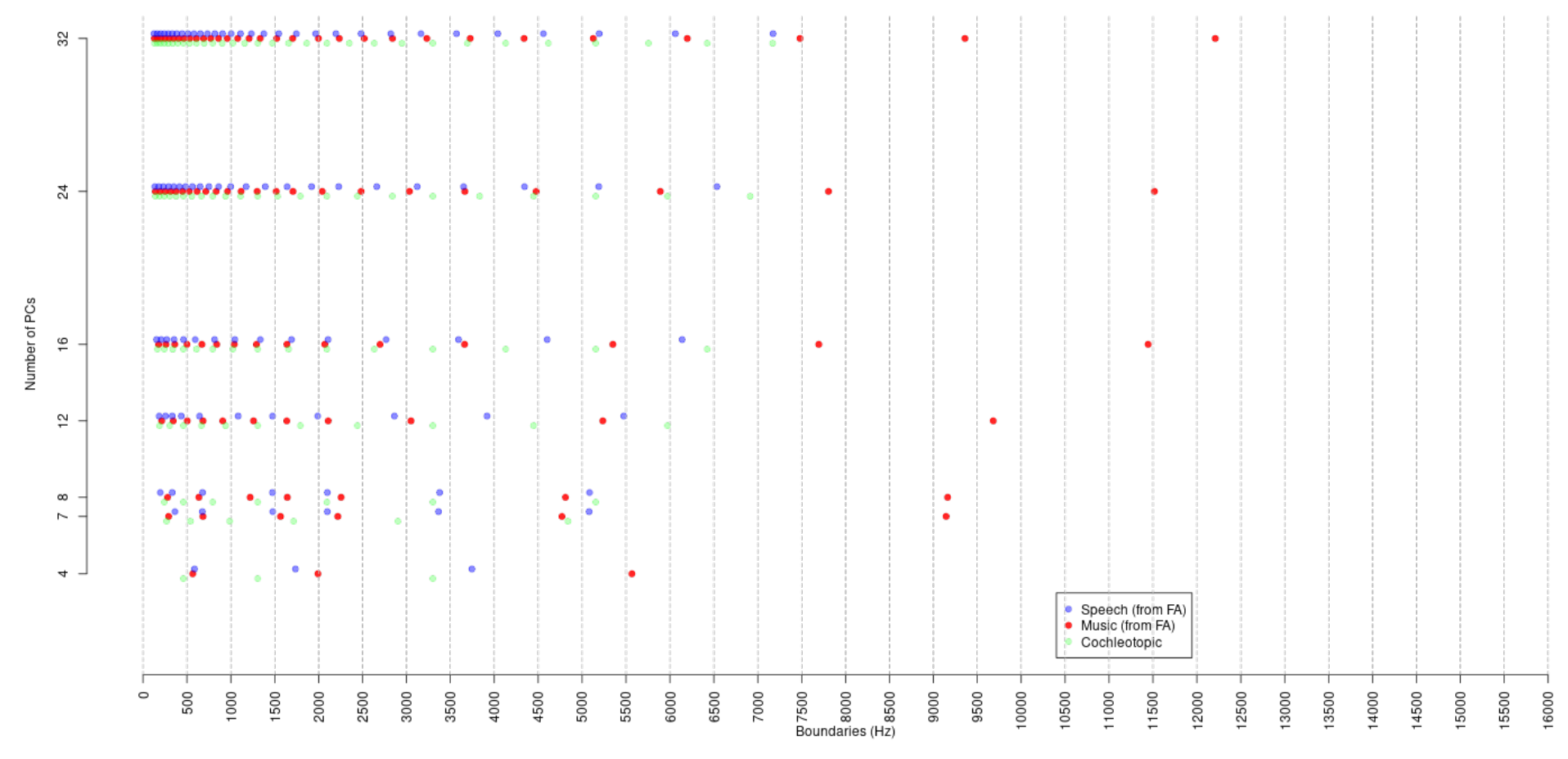

Figure 12: Boundary comparison (blue: speech boundaries from FA, red:music boundaries from FA, green: logarithmic boundaries)

PC Ranks comparison music vs speech, 16 kHz

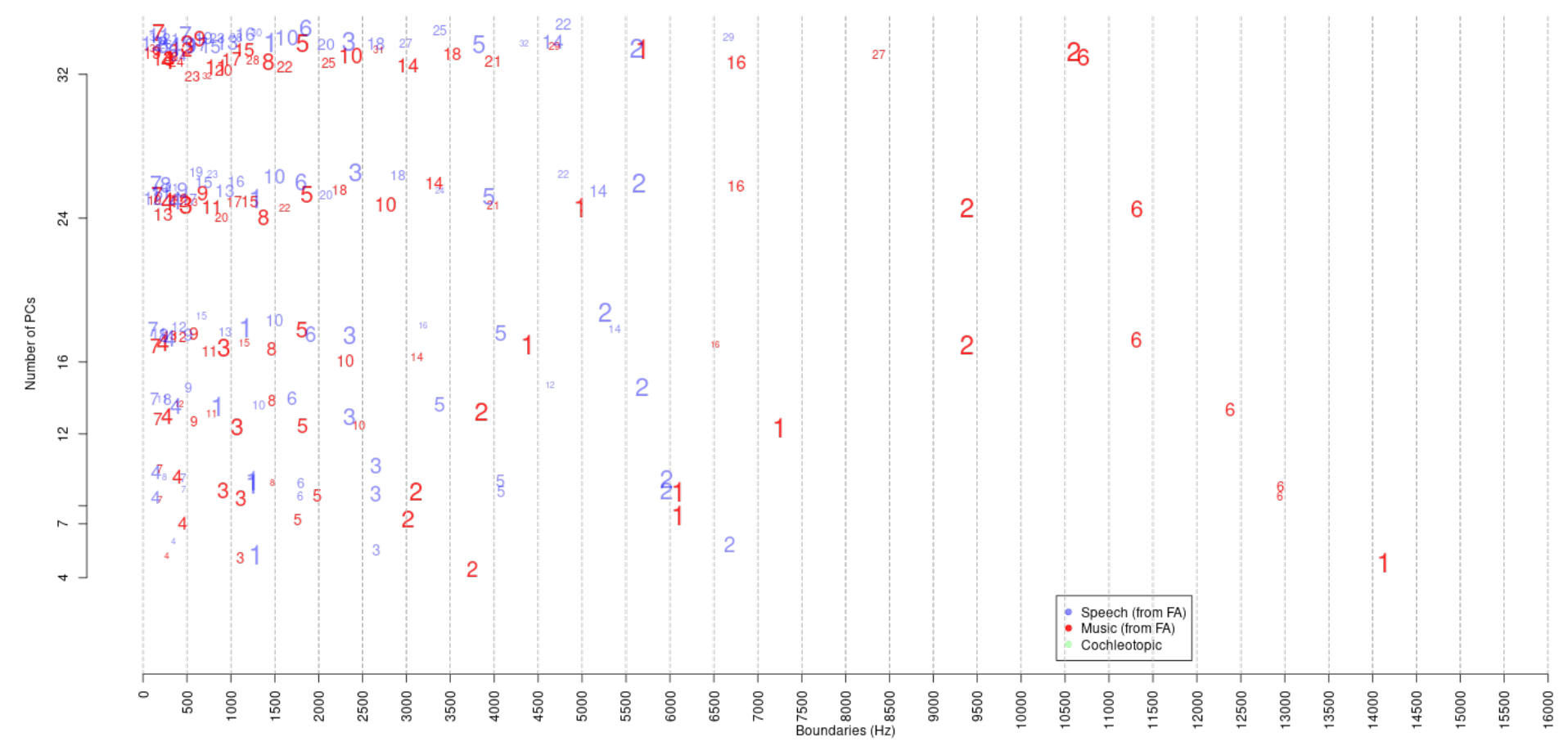

Figure 13: PC ranks comparison (blue: speech PC from FA, red:music PC from FA)

Nantes Université cnrs lling Région PAYS DE LA LOIRE RECHERCHE FORMATION INNOVATION en Pays de la Loire WEST CREATIVE INDUSTRIES Research, Education & Innovation in Pays de la Loire

agnieszka.duniec@etu.univ-nantes.fr