ASR Feature Extraction with Morphologically-Filtered Power-Normalized Cochleograms

F. de-la-Calle-Silos, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Peláez-Moreno
Universidad Carlos III de Madrid

Abstract

- Modeling of the **masking** behavior of the **Human Auditory System** to enhance the robustness of the feature extraction stage.
- Non-linear filtering of a **spectro-temporal representation** applied simultaneously on both the frequency and time domains, by processing it using mathematical morphology operations **as if it were an image**.
- A particularly important component of this architecture is the so called structuring element.
- On the Aurora 2 noisy continuous digits task, we report relative error reductions of 18.7% compared to PNCC and 39.5% compared to MFCC.

Morphological Filtering (MF)

Closing operator:

$$S \bullet M = (S \oplus M) \ominus M$$

Preserves the regions that have a similar shape as the structuring element.

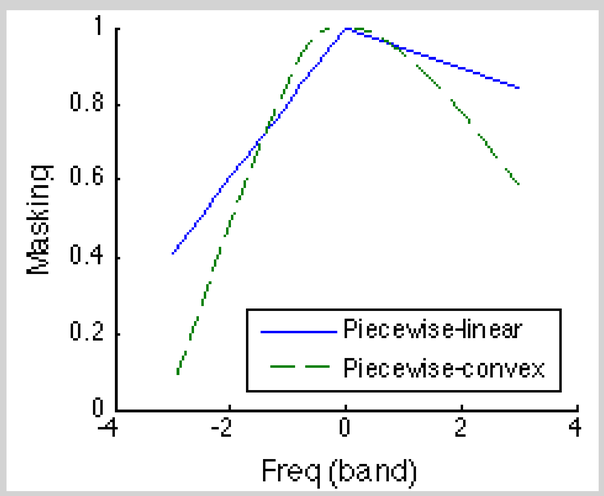
Masking: biological evidences

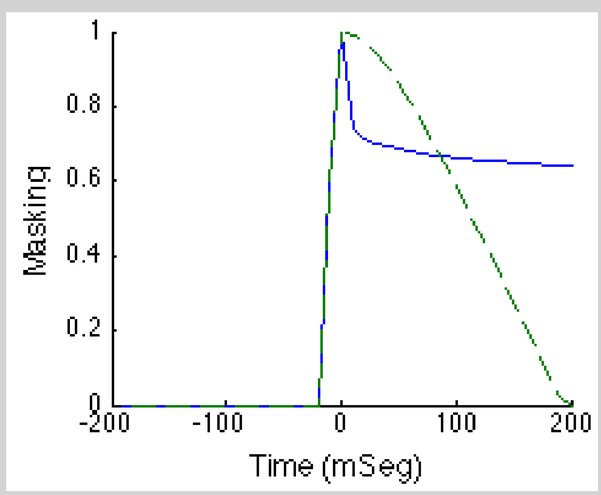
- Auditory masking occurs when the perception of one sound is affected by the presence of another sound.
- Auditory masking in the frequency domain is known as **simultaneous masking**, and in the time domain is known as **temporal masking**.
- The simultaneous masking is better represented in logarithmic scales where the spacing and the masker slope are regularly extended [4].
- The simultaneous masking is modeled with a slope of $+30dB$ per band for the lower bands and $-8db$ per band for the upper bands.
- The premasking effect is modeled as a constant slope of $+25dB/mSeg$,
- A fitted model for single masker-induced postmasking presented in [2] is used to model the postmasking effect:

$$M(t - T_m, L_m) = a(b - \log T_m)(L_m - c)$$

Structuring Element (SE): a spectro-temporal induction

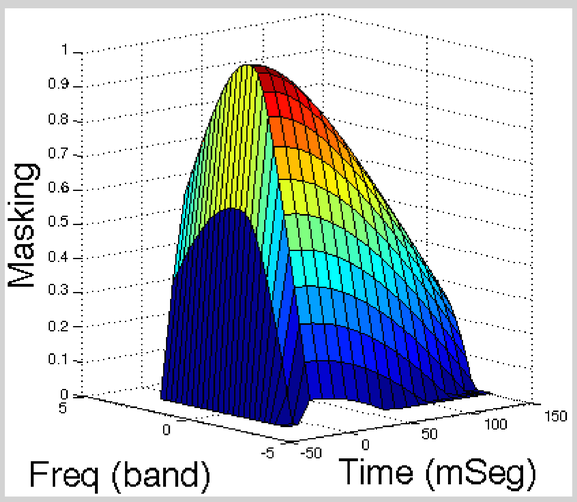
- Masking SE for a single frequency-time point is **very sharp**. [1] point a smooth model around (Fm, Tm)
- We hypothesize that the masking response of a particular (Fm, Tm) must be the **aggregation of many single-point responses**.
- Propose a **piecewise-convex model** built by aggregating 4 paraboloid quadrants of different parameters fitted to the contour provided by the linear model.

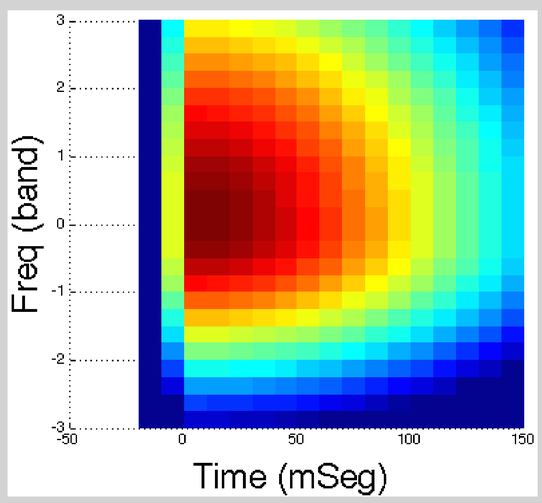




Comparison between the piecewise-linear model and the proposed piecewise-convex model.

Resulting structuring element



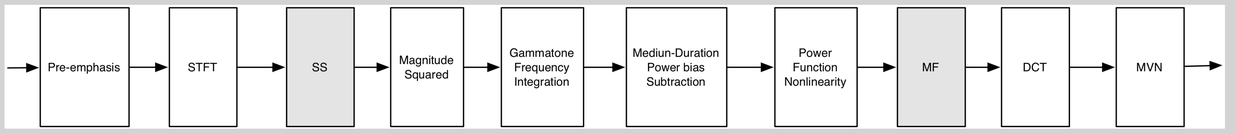


Diferents views of the structuring element, colour represents the weight of each pixel in the morphological operations.

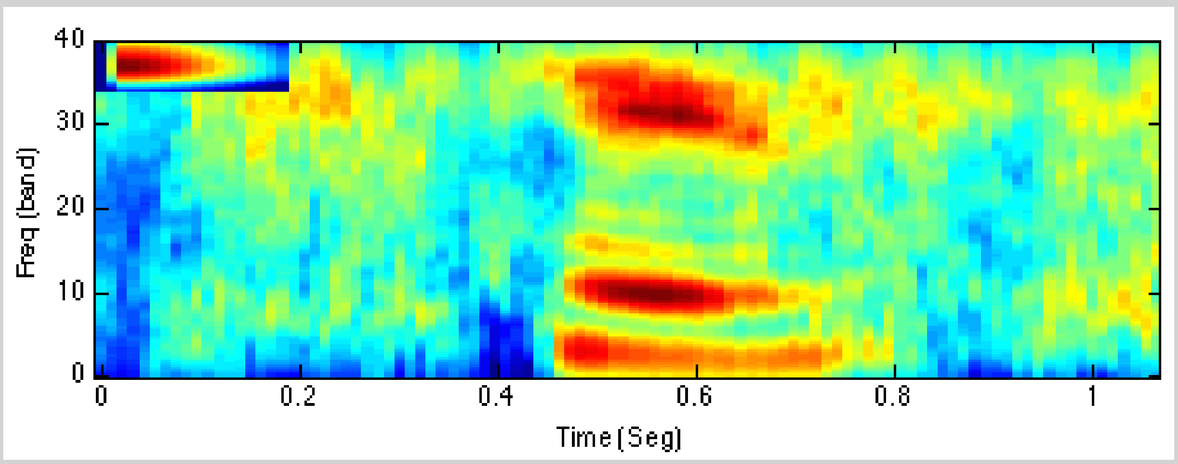
Underlying spectro-temporal representation

- Single mask across all frequencies and intensities despite the fact that the masking properties are **frequency and sound intensity-dependent** relying on the underlying spectro-temporal representation to accommodate these effects.
- We have chosen the power-normalized spectro-temporal representation used in the **Power-Normalized Cepstral Coefficients (PNCC)** [3] feature extraction process: Equivalent Rectangular Bandwidth (ERB) scale and a gammatone-shaped filter bank analysis

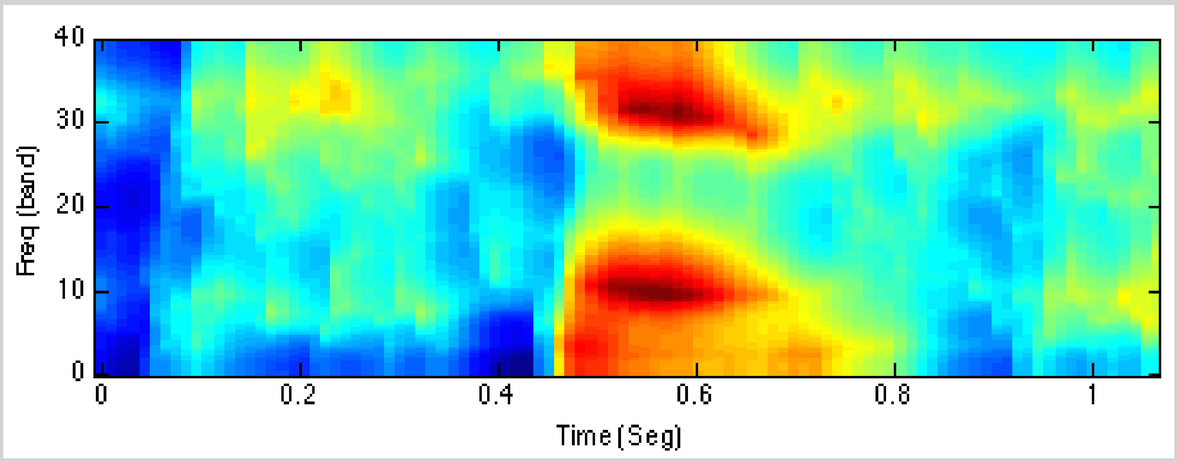
Proposed front-end algorithm



Examples of Spectrograms

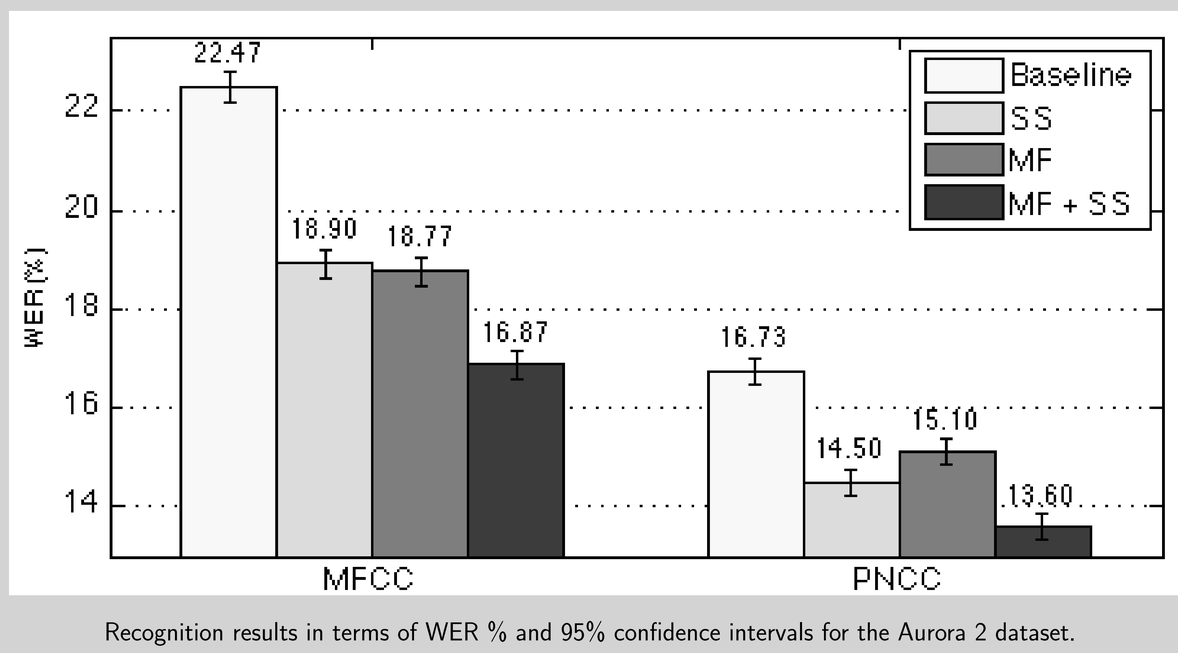


Noisy Spectrogram, in the upper left the structuring element can be appreciate.



Spectrogram aflter morphological filtering. This is the result of the operation: $S \bullet M$

Results



Recognition results in terms of WER % and 95% confidence intervals for the Aurora 2 dataset.

Conclusions

- Biologically-motivated SE that takes into **account the HAS masking properties** is presented.
- **Combination of PNCC** with spectral subtraction and morphological processing.
- A significant **increase in recognition rates** in the Aurora 2 dataset.

References

1. Hugo Fastl and Eberhard Zwicker. *Psycho-acoustics: Facts and Models*. 3rd ed. Springer, 2007.
2. W Jesteadt, S P Bacon, and Lehman JR. "Forward masking as a function of frequency, masker level, and signal delay." In: *The Journal of the Acoustical Society of America* (1982).
3. Chanwoo Kim and Richard M Stern. "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* ().
4. E. Zwicker and A. Jaroszewski. "Inverse frequency dependence of simultaneous tone-on-tone masking patterns at low levels". In: *The Journal of the Acoustical Society of America* (1982).