Paper and References

Efficient Auditory Coding in Speech Categorization

Supervisor: Dimme de Groot Edwin van der Heijden Professor: Jorge Martinez

EEMCS - Delft University of Technology

What are Auditory Kernels?

Auditory Kernels are the sparse, efficient representation of speech signals. The acoustic signal of speech is efficiently encoded like the human auditory system (the cochlea and cortex), according to the efficient coding hypothesis [1]. Every sound can be composed of spikes in time of these kernels, placed in time with amplitude.

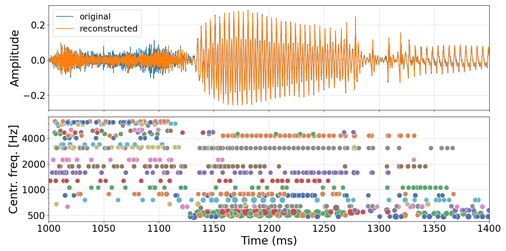


Figure 1. A snippet of a reconstruction of the utterance SA1 from the TIMIT corpus [8] using an auditory kernel set trained on TIMIT utterances (Figure 2c). Top: the original waveform (blue) overlaid with its reconstruction (orange). Bottom: the corresponding spike representation, where each dot marks a placed kernel, positioned by its time (x-axis) and the kernel's centroid frequency (y-axis). Each color identifies a unique kernel.

A speech signal can be represented by the following formula:

$$x(t) = \underbrace{\sum_{m=1}^{M} \sum_{i=1}^{n_m} s_i^m \, \phi_m(t - \tau_i^m)}_{\text{Reconstruction}} + \underbrace{\varepsilon^M(t)}_{\text{Residual}} \qquad (1)$$

where $\phi_1, \ldots, \phi_M$ are the kernel functions, τ_i^m is the temporal position, and s_i^m is the amplitude of the i-th spike of kernel ϕ_m, n_m is the number of spikes for that kernel, and $\varepsilon^M(t)$ is the residual not captured by the selected kernel dictionary of length M.

Phoneme Categorization

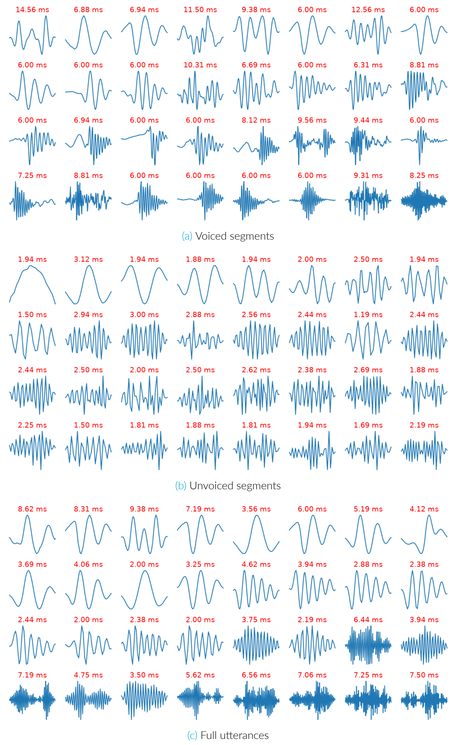
Category	Phonemes
Voiced	
Vowels	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr
Semivowels/Glides	l, r, w, y, el, hv
Nasals	m, n, ng, em, en, eng, nx
Fricatives	v, dh, z, zh
Affricates	jh
Stops	b, d, g, dx
Unvoiced	
Fricatives	f, th, s, sh
Affricates	ch
Stops	p, t, k, q
Glottal	hh
Stop Closures	bcl, dcl, gcl, pcl, tcl, kcl
Devoiced Schwa	ax-h

Table 1. Phonemes used in TIMIT divided by voicing, as indicated in the IPA [5]

Research Question

How do efficient auditory kernel sets learned from voiced versus unvoiced phonemes differ in their encoding efficiency, and do voiced and unvoiced kernels transfer differently to animal vocalizations versus environmental sounds, as the harmonic-transient distinction would predict?

Trained Kernel Sets



(a) Voiced segments

(b) Unvoiced segments

(c) Full utterances

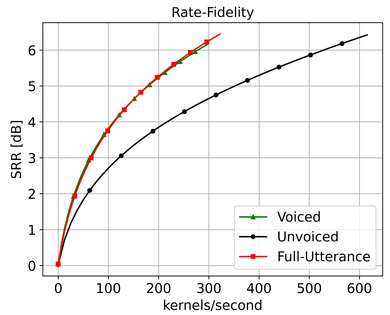
Figure 2. Three sets of 32 kernels, learned on voiced segments, unvoiced segments or full utterances of the TIMIT corpus, shown as time-domain waveforms and sorted by centroid frequency. The number above each kernel in red gives its length in milliseconds.

Phoneme Error Rates (PER)

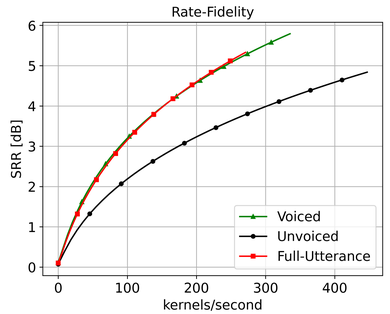
Kernel set	PER	Voiced PER	Unvoiced PER
Full-utterance reconstructions			
Baseline	0.0756	0.0833	0.0680
Voiced	0.1273	0.1356	0.1453
Unvoiced	0.1249	0.1347	0.1339
Full-Utterance	0.1267	0.1361	0.1403
Voiced reconstructions			
Voiced	0.2775	0.2595	0.1379
Unvoiced	0.2812	0.2624	0.1190
Full-Utterance	0.2789	0.2623	0.1131
Unvoiced reconstructions			
Voiced	0.7643	0.3033	0.5854
Unvoiced	0.7411	0.3234	0.5495
Full-Utterance	0.7695	0.3356	0.5795

Table 2. Average phoneme error rate (PER) [10] for each kernel set, grouped by reconstruction type (full utterances, voiced segments, unvoiced segments). The PER column scores all phonemes; the Voiced PER and Unvoiced PER columns score only phonemes of that voicing category in the reference. The baseline row is the original, non-reconstructed utterance, giving the recognizer's floor.

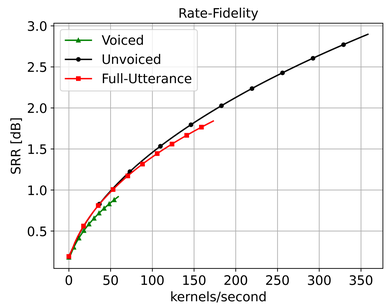
Rate-Fidelity: TIMIT



(a) Full utterances from the TIMIT test set



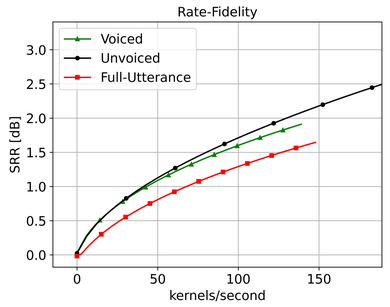
(b) Voiced segments



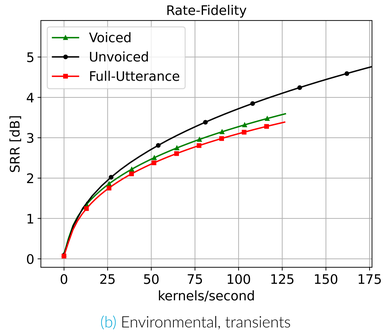
(c) Unvoiced segments

Figure 3. Rate-fidelity curves for reconstructing TIMIT speech with the voiced (green triangles), unvoiced (black dots), and full-utterance (red squares) kernel sets. Each panel plots signal-to-residual ratio (SRR) against kernels per second; a higher curve indicates more efficient reconstruction.

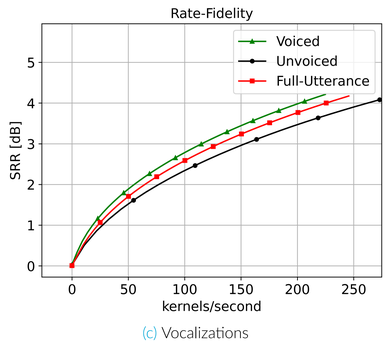
Rate-Fidelity: ESC-50



(a) Environmental, ambients



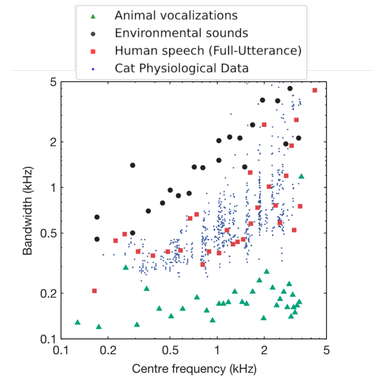
(b) Environmental, transients



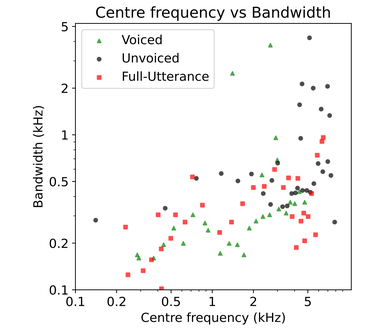
(c) Vocalizations

Figure 4. Rate-fidelity curves for reconstructing environmental sounds and vocalizations with the voiced (green triangles), unvoiced (black dots), and full-utterance (red squares) kernel sets. Each panel plots signal-to-residual ratio (SRR) against kernels per second; a higher curve indicates more efficient reconstruction.

Kernel Spectral Features



(a) Kernels trained by Smith and Lewicki on animal vocalizations (green triangles), environmental sounds (black dots), and human speech (red squares); the small blue dots are kernels derived from cat physiological data rather than trained. Adapted from [3].



(b) Our kernels trained on voiced segments (green triangles), unvoiced segments (black dots), and full utterances (red squares).

Figure 5. Centre frequency versus bandwidth of the trained kernels, comparing the sound-class kernels of Smith and Lewicki [3] (5a) with our phoneme-class kernels (5b). Panel (5a) is adapted from [3]

Conclusions and Future Work

- **Structure:** Kernels trained on voiced and unvoiced phonemes diverge clearly, voiced are long, narrow-band, and harmonic; unvoiced are short, broadband, and aperiodic.
- **Efficiency:** Each set reconstructs its own phoneme class most efficiently, while the full-utterance set spans both as a compromise code.
- **The mirror holds:** Voiced kernels reconstruct vocalizations best and unvoiced kernels transient sounds best.
- **Fidelity ≠ intelligibility:** The choice of kernel set has little effect on PER, even though rate-fidelity separates the sets clearly.
- **Limitations:** PER captures phoneme-level transcription accuracy rather than whether the message would be understood; PER is coarse over short segments; kernel-length ranges and training durations were not equalized.
- **Future Work:** Train kernels on transience rather than voicing as the organizing axis; analyze per-segment kernel usage in the full-utterance set for voicing categories; test whether unvoiced kernels capture broadband noise in general.