On the Information Rate of Speech Communication

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI VICTORIA UNIVERSITY OF WELLINGTON

STEVEN VAN KUYK[Vict], W. BASTIAAN KLEIJN[Vict,TUD] AND RICHARD C. HENDRIKS[TUD]

TU Delft Delft University of Technology

Problem

- The success of speech-based technology relies on an understanding of speech communication.
- Common approaches:
  - linguistic models (phonemes, syllables, words, sentences).
  - speech production models (source-filter theory, articulatory phonetics).
  - auditory models (cochlear filtering, equal loudness curves, hair cell transduction).
- Needed: a model of human speech communication derived from first principles, i.e., information theory.

Contributions

- An upper bound on the information rate of speech of the order of 100 b/s.
- A model of speech communication that does not rely on prior knowledge of the transmitter (vocal tract), the receiver (auditory system), or language.
- The model only relies on having recordings of multiple talkers saying the same utterance.

Previous work

- The probabilities of English phonemes give a lexical information rate of approximately 50 b/s [Fano, 1950].
- Variables related to talker identification, emotional state, and prosody, vary relatively slowly in time and contribute little to the overall information rate.
- The bandwidth of the human auditory system and the SNR required for perfect intelligibility give a channel capacity of approximately 20000 b/s [Flanagan, 1972].
- Why are information rates based on acoustics orders of magnitude larger than those based on linguistics?
- Hypothesized that talker variability is a type of 'production noise' that limits information transfer [Kleijn *et al.*, 2015].

Human communication

- A talker randomly selects a message, $\{\mathbf{M}_t\}$, e.g., a phoneme, word, or neural state, where t is the time index.
- The talker encodes the message into an acoustic speech signal, $\{\mathbf{S}_t\}$, according to a conditional probability distribution:

$$p_{\{\mathbf{S}_t\}|\{\mathbf{M}_t\}}(\{S_t\}|\{M_t\}).$$

- Define a *chorus* as a set of J speech signals where each signal contains the same message:

$$\{\mathbf{Z}_M\} = \{\{\mathbf{S}_{M,t}\}^{(1)}, \{\mathbf{S}_{M,t}\}^{(2)}, \cdots, \{\mathbf{S}_{M,t}\}^{(J)}\},$$

- Define a chorus-based estimate of the message as $\{\tilde{\mathbf{M}}_t\} = f(\{\mathbf{Z}_M\}) + \{\mathbf{N}_t\}$ where $f(\cdot)$ is a deterministic function, $\{\mathbf{N}_t\}$ is regularization noise.

Information bottleneck

- A natural objective for the estimator $f(\cdot)$ is that it minimizes the information bottleneck:

$$f^* = \arg\min_f \; I(\{\mathbf{Z}_M\}; \{\tilde{\mathbf{M}}_t\}) - \beta I(\{\mathbf{S}_{M,t}\}; \{\tilde{\mathbf{M}}_t\}), \quad (1)$$

- $I(\{\mathbf{Z}_M\}; \{\tilde{\mathbf{M}}_t\})$ is the mutual information rate between the chorus and the chorus message estimate.
- $I(\{\mathbf{S}_{M,t}\}; \{\tilde{\mathbf{M}}_t\})$ is the mutual information rate between the speech and the message estimate.
- β is a Lagrange multiplier.
- Main idea: the bottleneck discards features from the chorus that aren't consistent across talkers.

Comparison of estimators

- Confine the message estimator to be of the form $f(\{\mathbf{Z}_M\}) = \frac{1}{J}\sum_j g(\{\mathbf{S}_{M,t}\}^{(j)})$.
- Consider $g(\cdot)$ as the identity function, STFT, spectrogram, log-spectrogram, and auditory spectrogram. Of these candidate functions, the auditory spectrogram gives the lowest bottleneck. Hence, we represent speech as a sequence of auditory-spectra: $\{\mathbf{X}_t\} = g(\{\mathbf{S}_{M,t}\})$.
- Suggests that the structure of speech might be adapted to the coding capability of the mammalian auditory system.

The information rate

- Describe speech communication by

$$\{\mathbf{X}_t\} = \{\tilde{\mathbf{M}}_t\} + \{\tilde{\mathbf{P}}_t\}, \quad (2)$$

$\mathbf{X}_t$ is the auditory-spectra of the speech, $\tilde{\mathbf{M}}_t$ the estimated message, $\tilde{\mathbf{P}}_t$ is Gaussian production noise.
- The mutual information rate is

$$I(\{\mathbf{X}_t\}; \{\tilde{\mathbf{M}}_t\}) = \lim_{k\to\infty} \frac{1}{k} I(\mathbf{X}^k; \tilde{\mathbf{M}}^k), \quad (3)$$

where $\mathbf{X}^k$ and $\mathbf{M}^k$ are formed by stacking k consecutive spectra. This means that time and frequency dependencies are accounted for.
- If time-dependencies span no more than L samples, i.e., $\mathbf{X}_t$ is independent to $\mathbf{X}_{t+L}$, then the mutual information rate reduces to

$$I(\{\mathbf{X}_t\}; \{\tilde{\mathbf{M}}_t\}) = \frac{1}{L}\left(h(\mathbf{X}^L) - h(\tilde{\mathbf{P}}^L)\right), \quad (4)$$

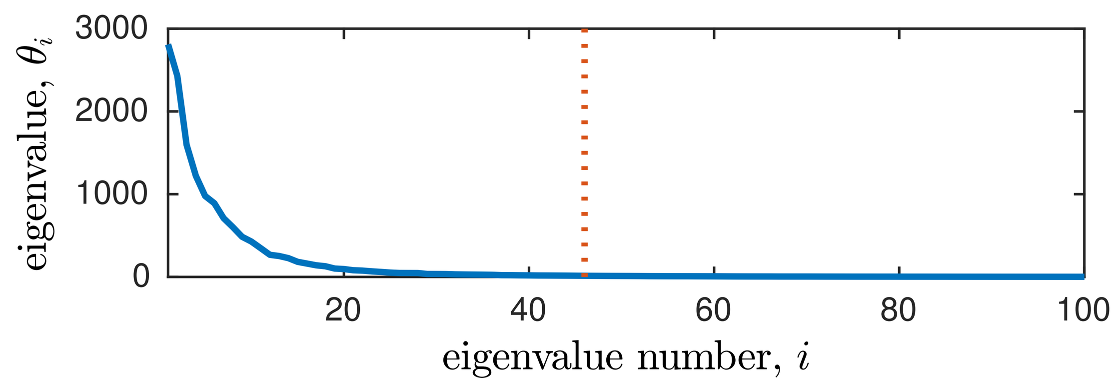
- Use Principal Component Analysis to find a subspace containing the message estimate (called the *message articulation space*).
- Reduce dimensionality by projecting stacked spectra onto the message articulation space.
- The capacity of the speech communication channel is

$$C = \frac{F}{2L}\sum_{v=1}^{V} \log_2 \frac{\lambda_v}{\psi_v}, \quad (5)$$

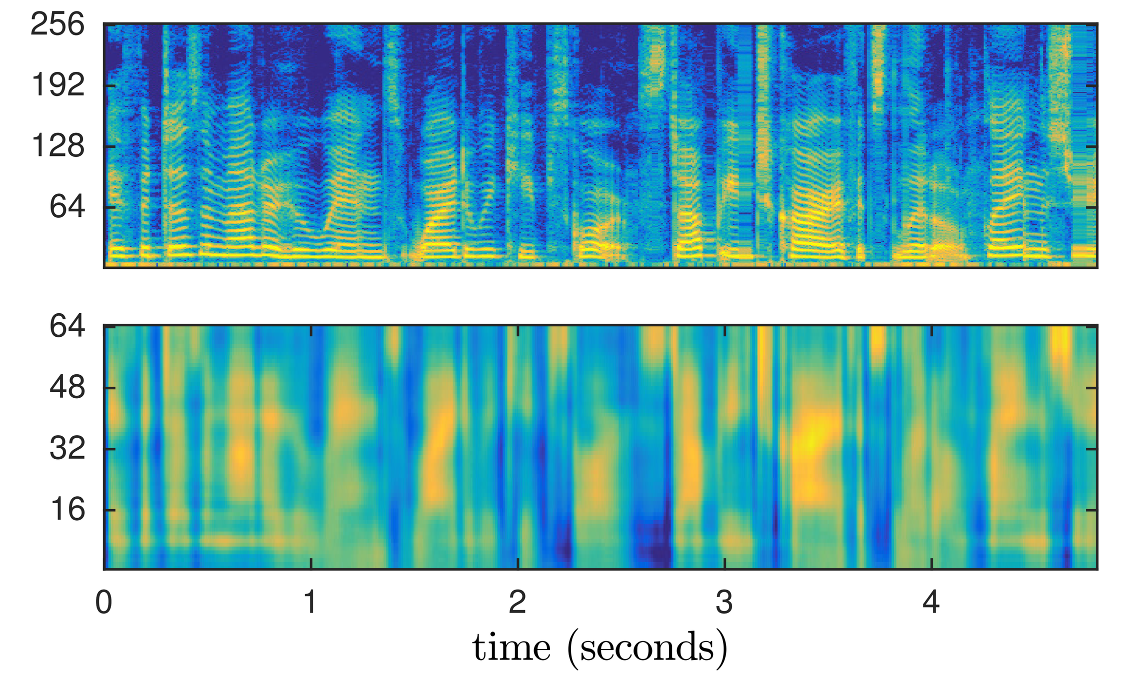
where F is the frame rate and λ_v and ψ_v are the eigenvalues of the covariance matrices of the speech and the production noise after projecting onto the message articulation space.

Message articulation space

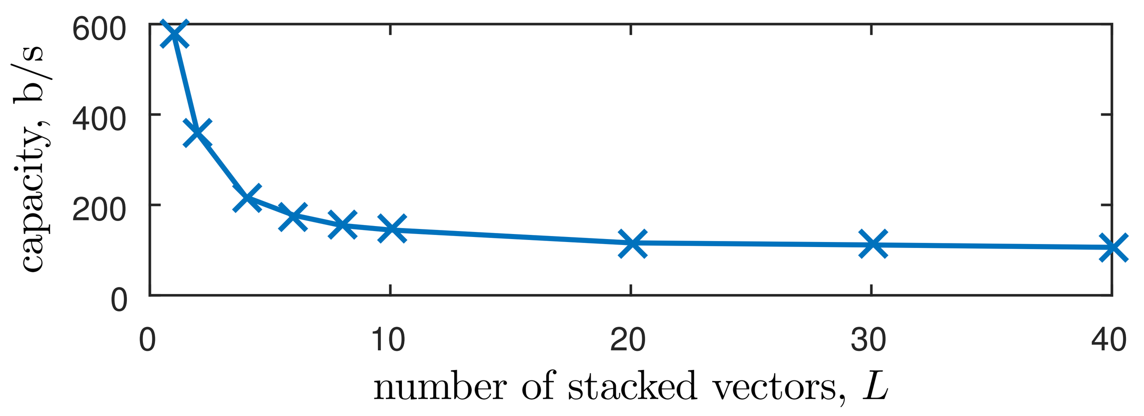
- The message estimate lies within an approximately 50-dimensional subspace:



Projected speech spectra



Channel capacity



Results

- The capacity of the speech communication channel is of the order of 100 b/s, which is comparable to the lexical information rate.
- Time-dependencies are negligible for $L > 10$. This corresponds to a duration of 80 ms, which is consistent with the average duration of a phoneme.

Discussion

- Phonemes and words are not closed under addition. In reality the message articulation space is a manifold, not a subspace. This would lead to a lower information rate.
- In theory, given enough data, our model can also account for phoneme and word dependencies.
- Of all the representations of speech we tried, the auditory-spectra gave the lowest bottleneck, but there could be another representation that achieves a lower bottleneck. Such a representation could be found by solving (1).