ON THE INFORMATION RATE OF SPEECH COMMUNICATION

Steven Van Kuyk[1], W. Bastiaan Kleijn[1,2] and Richard C. Hendriks[2]

[1]Victoria University of Wellington, New Zealand
[2] Delft University of Technology, The Netherlands

ABSTRACT

The key to the success of speech-based technology is an understanding of human speech communication. While significant advances have been made, a unified theory of speech communication that is both comprehensive and quantitative is yet to emerge. In this paper we approach speech communication from an information theoretical perspective. Without relying on prior knowledge of speech production, language, or auditory processing, we develop a new methodology for measuring the information rate of speech. Instead we rely on having recordings of multiple talkers saying the same utterance. In general, our results are consistent with a linguistic understanding of speech communication.

Index Terms— Information theory, speech communication, talker variability

1. INTRODUCTION

Shannon's information theory (IT) [1] provides a mathematical framework for analyzing communication systems, regardless of the systems implementation. There are two key concepts to IT. The first concept is that the meaning of a message is irrelevant to the engineering problem. Rather, the significant aspect is that a transmitted message is one selected from a set of possible messages. This view leads to a probabilistic approach. The second concept of IT is that a message of low probability contains more information than a message of high probability. IT is fundamental to the design of wireless communications, cryptography, and data compression systems [2]. There is no reason that IT cannot be applied to models of human speech communication. Surprisingly, a relatively small effort has been made to do so.

In the context of speech communication the most fundamental question to ask is ''What is the maximum amount of information that can be transferred from a talker to a listener per unit of time''? The answer to this question is important because it gives a criterion for assessing the effectiveness of speech-based communication systems. Furthermore, it could suggest new algorithms for speech enhancement, speech coding, and speech recognition.

Broadly speaking, two approaches to measuring the information rate of speech exist: the linguistic approach, and the acoustic approach. The linguistic approach describes speech as a sequence of discrete perceptual units such as phonemes, words, or sentences. Taking the average talking speed as 12 phonemes per second [3], and using the English phoneme probabilities tabulated in [4], the lexical information rate is approximately 50 b/s. When the dependencies of the phonemes are accounted for the rate will be decreased further.

The lexical information rate does not include information about talker identification, emotional state, and prosody. However, these variables vary relatively slowly in time and contribute little to the overall information rate. As an example, [5] estimated that the total amount of talker-specific information (e.g., age, accent, sex) was of the order of 30 bits. If speech with a duration of one minute is considered, then accounting for talker-specific information increases the information rate by only 0.5 b/s.

The acoustic approach to measuring the information rate describes speech communication using statistical models of acoustic signals. By considering the bandwidth of the human auditory system and the speech-to-noise ratio required for perfect intelligibility, [6] concluded that the speech communication channel could support an information rate of 20000 b/s. In contrast, by considering a simple model of speech production, [5] estimated that the information rate was of the order of 1500 b/s.

Why do information rates based on acoustic speech signals tend to be orders of magnitude larger than estimates based on linguistics? In [5], Fano hypothesized that talker variability behaves as a type of internal noise that limits the effectiveness of communication. Under this point of view, each talker encodes lexical information into acoustic signals in a unique way. This idea was formalized in [7], which coined the term 'production noise' and showed theoretically that the usefulness of a communication channel must saturate either at the 'message-to-production noise-ratio' or at the speech-to-enviromental-noise ratio, whichever is lower. The resulting relation in [7] between intelligibility and environmental noise spectra closely resembles that of the heuristically derived measures such as the articulation index (AI), e.g., [8, 9], and the speech intelligibility index (SII) [10, 11].

In this paper, we present a new method for measuring the information rate of speech that is based on the effect of talker variability. Unlike existing methodologies, our approach does not require knowledge of language, speech production, or the human auditory system. Instead we rely on data in which different talkers speak the same utterance (referred to as a "chorus"). We generalize the speech communication model proposed in [7] to include time-frequency dependencies. To locate the relevant information encoded in acoustic speech signals we use the information bottleneck principle [12, 13]. Thus, we find an upper bound on the rate in terms of a capacity of a channel in the feature space found with the bottleneck. This upper bound is around 100 b/s.

The remainder of this paper is organised as follows. In the following section we describe an information theoretical model of speech communication. Section 3 uses the model to measure the information rate of speech and Section 4 concludes the work.

2. THEORY

The basic concept of our approach is to extract the information that is consistent between talkers who speak the same utterance. To that purpose we use a *chorus* of talkers. This enables us to distinguish the production noise and the actual message. To make the approach

work in practice, we must use a suitable representation of the signal. We find that representation using the information bottleneck.

In the following, we denote all random variables with bold font and their realizations in non-bold font. We characterize speech signals and the underlying message as stationary discrete-time random processes. For example, the message process is written as $\{\mathbf{M}_t\} = \{\mathbf{M}_t \mid t \in \mathbb{Z}\}$ where t is the time index.

2.1. Model of speech communication

Let us consider the nature of speech communication. First, a talker randomly selects a message for transmission according to a probability distribution $p_{\{\mathbf{M}_t\}}(\{M_t\})$ where $\{\mathbf{M}_t\}$ is the message. While this is not part of our formalism, which is not based on linguistics, the message may be thought to represent a sequence of phonemes, words, sentences, or neural states.

The talker encodes the message into an acoustic speech signal, $\{\mathbf{S}_t\}$, according to a conditional probability distribution $p_{\{\mathbf{S}_t\}|\{\mathbf{M}_t\}}(\{S_t\}|\{M_t\})$. In this way, the variability of speech produced by different talkers is incorporated into the model.

We define a random chorus as a set of J speech signals $\{\mathbf{Z}_M\} = \{\{\mathbf{S}_{M,t}\}^{(1)}, \{\mathbf{S}_{M,t}\}^{(2)}, \cdots, \{\mathbf{S}_{M,t}\}^{(J)}\}$ where each speech signal in the set contains the same message. Here, the subscript M indicates a particular message. A chorus-based estimate of the message is defined by $\{\tilde{\mathbf{M}}_t\} = f(\{\mathbf{Z}_M\}) + \{\mathbf{N}_t\}$, where $f(\cdot)$ is a deterministic function and $\{\mathbf{N}_t\}$ is a small amount of multivariate Gaussian noise that is statistically independent to $\{\mathbf{Z}_M\}$. This regularization noise corresponds to the situation that the message is estimated from $\{\mathbf{Z}_M\}$ with some uncertainty, and ensures that the mutual information between the chorus and the message estimate is bounded from above. Depending on the estimator $f(.)$, $\{\tilde{\mathbf{M}}_t\}$ could lie in a different domain to $\{\mathbf{M}_t\}$. However, an ideal estimator will result in a one-to-one relationship between the domain of $\{\tilde{\mathbf{M}}_t\}$ and the domain of $\{\mathbf{M}_t\}$. We call the domain of $\{\tilde{\mathbf{M}}_t\}$ the *message articulation space* (MAS). Signal components that lie outside of the MAS are considered irrelevant to the communication process.

2.2. The information bottleneck

A natural objective for $f(\cdot)$ is that it minimizes the information bottleneck [12, 13]

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

where $I(\{\mathbf{Z}_M\}; \{\tilde{\mathbf{M}}_t\})$ 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, and β is a Lagrange multiplier.

On the one hand, optimizing the information bottleneck leads to an operator $f(\cdot)$ that creates a compressed description of the chorus as it minimizes $I(\{\mathbf{Z}_M\}; \{\tilde{\mathbf{M}}_t\}))$. Minimizing this term gives the estimator a disincentive to simply accumulate in the message estimate all speech signals in the chorus, or even to select a single speech signal as the message estimate. On the other hand, optimizing the bottleneck maximizes the information shared between the message estimate and speech carrying the same message. Note that the mutual information rate $I(\{\mathbf{S}_{M,t}\}; \{\tilde{\mathbf{M}}_t\})$ cannot exceed the true message information rate in the signal as the speech and the chorus used in the estimator are independently drawn from $p_{\{\mathbf{S}_t\}|\{\mathbf{M}_t\}}(\{S_t\}|\{M_t\})$. Thus, overweighting this term ($\beta \gg 1$) will not result in an increase in its value. Such overweighting of the second term also prevents that the optimal estimator would be the trivial estimator that always maps to zero.

In practice it is difficult to evaluate (1) due to computational complexity. However, given a finite set of candidate estimators $f(\cdot)$ we can evaluate the information bottleneck for each $f(\cdot)$ and select the $f(\cdot)$ that achieves the lowest meaningful bottleneck. In this way the information bottleneck acts as a criterion for evaluating how successful $f(\cdot)$ is at estimating the message conveyed by the chorus.

We confine the estimator $f(\cdot)$ to be of the form

$$f(\{\mathbf{Z}_M\}) = \frac{1}{J}\sum_j g(\{\mathbf{S}_{M,t}\}^{(j)}), \quad (2)$$

where $g(\cdot)$ is a surjective mapping onto the MAS. Note that this form indeed can not store all speech in the chorus.

2.3. Comparison of message estimators

We now provide some examples for the mappings $g(\cdot)$ from the acoustic signal representation. First we consider the case where $g(\cdot)$ is the identity function. Then, asymptotically with increasing J, we have

$$\begin{aligned}\lim_{J\to\infty}\{\tilde{\mathbf{M}}_t\} &= \{\mathbf{N}_t\} + \lim_{J\to\infty}\frac{1}{J}\sum_j \{\mathbf{S}_{M,t}\}^{(j)} \\ &= \{\mathbf{N}_t\} + \{\mathrm{E}[\mathbf{S}_M]_t\} \\ &= \{\mathbf{N}_t\} + \{0_t\},\end{aligned} \quad (3)$$

where $\mathrm{E}[\cdot]$ is the expectation operator. The third equality follows because acoustic waveforms produced by different talkers are statistically independent. Since the regularization noise is drawn independently, we have that asymptotically $I(\{\mathbf{Z}_M\}; \{\mathbf{N}_t\}) = 0$ and that $I(\{\mathbf{S}_{M,t}\}; \{\mathbf{N}_t\}) = 0$, and hence the information bottleneck is zero.

As a second example, we consider the case where $g(\cdot)$ is the short-time Fourier transform (STFT) denoted $g(\{\mathbf{S}_{M,t}\}) = \{\mathbf{S}'_M(\omega)_t\}$, where ω is the frequency index. Then we have,

$$\begin{aligned}\lim_{J\to\infty}\{\tilde{\mathbf{M}}_t\} &= \{\mathbf{N}_t\} + \lim_{J\to\infty}\frac{1}{J}\sum_j \{\mathbf{S}'_M(\omega)_t\}^{(j)} \\ &= \{\mathbf{N}_t\} + \{\mathrm{E}[|\mathbf{S}'_M(\omega)|]\mathrm{E}[e^{j\angle\mathbf{S}'_M(\omega)}]_t\} \\ &= \{\mathbf{N}_t\} + \{\mathrm{E}[|\mathbf{S}'_M(\omega)|]\cdot 0_t\} \\ &= \{\mathbf{N}_t\} + \{0_t\},\end{aligned} \quad (4)$$

where we modeled the phase of the STFT of a speech signal as a uniformly distributed random variable between $-\pi$ and π that is statistically independent to the magnitude [14]. Similarly to the identity function, the STFT results in an information bottleneck equal to zero.

As a third example, we consider the case where $g(\cdot)$ is the spectrogram transform defined as the squared magnitude of the STFT. We denote the spectrogram as $g(\{\mathbf{S}_{M,t}\}) = \{\mathbf{X}'_{M,t}\}$. Here, $\mathbf{X}'_{M,t}$ is a vector where each element describes the power at a particular frequency location. Then,

$$\begin{aligned}\lim_{J\to\infty}\{\tilde{\mathbf{M}}_t\} &= \{\mathbf{N}_t\} + \lim_{J\to\infty}\frac{1}{J}\sum_j \{\mathbf{X}'_{M.t}\}^{(j)} \\ &= \{\mathbf{N}_t\} + \{\mathrm{E}[\mathbf{X}'_M]_t\} \\ &\approx \{\mathrm{E}[\mathbf{X}'_M]_t\},\end{aligned} \quad (5)$$

which results in a non-zero time-varying process that depends on an underlying message. In this case both mutual information rate terms

in the bottleneck are non-zero. Hence, by the overweighting β, the bottleneck will have a negative value.

As a fourth example, consider the logarithm of the spectrogram. In practice we have found that it results in a more negative information bottleneck than the spectrogram. The speech-production perspective provides an explanation: in the logarithmic domain the excitation information in the speech signal ends up in a separate, additive term that averages to zero across talkers [15].

Our approach to finding the information rate of speech does not require knowledge about speech production or the human auditory system: our search is for a representation that minimizes the bottleneck. This does not preclude us from evaluating a representation inspired by auditory models as an educated guess:

$$g(\{\mathbf{S}_{M,t}\}) = \{\log(F\mathbf{X}'_M)_t\}, \tag{6}$$

where F is a matrix that represents an auditory filterbank and the logarithm is applied elementwise. We set each row of F to be the squared magnitude response of a gammatone filter with the center frequency and bandwidth set according to the equivalent rectangular bandwidth scale (ERB) [16]. Of all the examples so far, we found that (6) gave the lowest bottleneck. Hence, we will use the auditory-spectra defined by (6) as our representation of speech. The fact that (6) gives a low bottleneck supports the results of [17], which suggested that the acoustic structure of speech might be adapted to the coding capability of the mammalian auditory system.

2.4. Computing the information rate of speech

We now derive the channel capacity for the feature space developed in the previous section. We model speech as a multi-dimensional ergodic stationary discrete-time random process that is described by

$$\{\mathbf{X}_t\} = \{\tilde{\mathbf{M}}_t\} + \{\tilde{\mathbf{P}}_t\}, \tag{7}$$

where $\mathbf{X}_t, \tilde{\mathbf{M}}_t, \tilde{\mathbf{P}}_t \in \mathbb{R}^N$ are column vector random variables that represent auditory-spectra at time $t \in \mathbb{Z}$. Herein $\mathbf{X}_t$ is the auditory-spectra of the speech, $\tilde{\mathbf{M}}_t$ the estimated message, and $\tilde{\mathbf{P}}_t$ the estimated production noise. Production noise was first introduced in [7] and has since been used to incorporate the effect of talker variability in speech intelligibility prediction [15] and speech enhancement [18, 19]. We assume that $\tilde{\mathbf{M}}_t$ and $\tilde{\mathbf{P}}_t$ are statistically independent and that $\tilde{\mathbf{P}}_t$ is zero-mean and multivariate Gaussian.

The mutual information rate between the speech and the message describes the effectiveness of communication. For vector processes consisting of a sequence of speech vectors, $\{\mathbf{X}_t\}$, and a sequence of estimated message vectors $\{\tilde{\mathbf{M}}_t\}$ the mutual information rate is

$$I(\{\mathbf{X}_t\}; \{\tilde{\mathbf{M}}_t\}) = \lim_{k\to\infty} \frac{1}{k} I(\mathbf{X}^k; \tilde{\mathbf{M}}^k), \tag{8}$$

where $\mathbf{X}^k = [(\mathbf{X}_1)^T, (\mathbf{X}_2)^T, \cdots, (\mathbf{X}_k)^T]^T$,
$\tilde{\mathbf{M}}^k = [(\tilde{\mathbf{M}}_1)^T, (\tilde{\mathbf{M}}_2)^T, \cdots, (\tilde{\mathbf{M}}_k)^T]^T$, T denotes the transpose, and $I(\mathbf{X}^k; \tilde{\mathbf{M}}^k)$ is the mutual information. Note that $\mathbf{X}^k$ is a $kN \times 1$ random vector obtained by stacking k consecutive speech vectors and similarly for $\tilde{\mathbf{M}}^k$.

The advantage of using the mutual information rate rather than the mutual information is that time-dependencies between successive spectra 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

$$\begin{aligned} I(\{\mathbf{X}_t\}; \{\tilde{\mathbf{M}}_t\}) &= \frac{1}{L} I(\mathbf{X}^L; \tilde{\mathbf{M}}^L) \\ &= \frac{1}{L}\big(h(\mathbf{X}^L) - h(\tilde{\mathbf{P}}^L)\big), \end{aligned} \tag{9}$$

where $h(\cdot)$ denotes differential entropy.

We recall that signal components that do not lie in the MAS do not contribute to the information rate. Exploiting this fact, we perform a dimensionality reduction by applying principal component analysis (PCA) to the stacked speech vectors. Given the relative scaling of the various dimensions, PCA uses a mean-square error criterion to discard dimensions that contribute the least variance. This is particularly reasonable for the Gaussian case because mutual information and differential entropy depend only on second-order statistics. Let R_{M^L} denote the covariance matrix of $\tilde{\mathbf{M}}^L$. An orthogonal basis for the MAS can be obtained from an eigen-decomposition:

$$R_{M^L} = U\Lambda U^T, \tag{10}$$

where the columns of U are the unit-magnitude eigenvectors of R_{M^L} and Λ is a diagonal matrix of the corresponding eigenvalues. Dimensionality reduction is then performed by removing eigenvectors from U with small eigenvalues. An orthogonal projection matrix that projects vectors onto the MAS is then given by

$$A = UU^T. \tag{11}$$

We can now define the channel capacity. It is defined as the maximum mutual information rate over all possible probability distributions of $\tilde{\mathbf{M}}^L$. For the multivariate Gaussian additive noise case with time-correlations no longer than L samples, we have that

$$C = \frac{F}{2L} \sum_{v=1}^{V} \log_2 \frac{\lambda_v}{\psi_v}, \tag{12}$$

where F is the sample rate of the process in Hz, V is the number of eigenvectors in U, λ_v is the v'th largest eigenvalue of R_{X^L} and ψ_v is the v'th largest eigenvalue of R_{P^L}, where R_{X^L} is the covariance matrix of $\mathbf{X}^L$ after projecting the stacked speech vectors onto the MAS, and R_{P^L} is the covariance matrix of $\tilde{\mathbf{P}}^L$ after projecting the stacked production noise vectors onto the MAS. The above equation can be obtained by first noting that the multivariate Gaussian distribution is the maximum entropy distribution for a given R_{X^L}, and then substituting the expressions for the entropy of multivariate Gaussian distributions into (9).

3. EXPERIMENTS

We now describe our experiment for measuring the capacity of the speech communication channel. First we describe our implementation and then we present our results.

3.1. Implementation

For our experiment, we create a chorus using data from the CHAINS speech corpus [20]. The CHAINS speech corpus includes easy reading material in English spoken by $J = 36$ talkers including 18 females, and 18 males. All speech signals were downsampled to a sampling rate of $f_s = 16$ kHz and normalized to have unit variance. Additionally, a dynamic time-warping algorithm [21] was applied to all signals in the chorus to ensure that the signals contained the same lexical information at a given time.

A sequence of auditory-spectra was computed for each signal in the chorus using a STFT with a 512-point Hann analysis window and 75% overlap. This gives a frame rate of $F = 125$ Hz. A gammatone filterbank that included 64 filters linearly spaced on the ERB-rate scale was then applied. We denote the sequence of spectra for the j'th talker $\{X_t\}^{(j)}$.

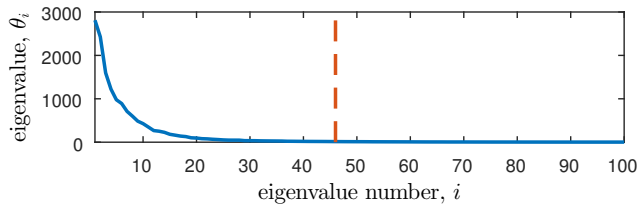

Fig. 1. The 100 largest eigenvalues of $\hat{R}_{ML}$. The red line at $V = 0.005\lambda_{\max}$ indicates the number of eigenvectors used.

The message and the production noise were estimated according to

$$\{\tilde{M}_t\} = \frac{1}{J}\sum_j \{X_t\}^{(j)} \tag{13}$$

and

$$\{\tilde{P}_t\}^{(j)} = \{X_t\}^{(j)} - \frac{1}{J-1}\sum_{l,l\neq j} \{X_t\}^{(l)}, \tag{14}$$

respectively. The message estimate (13) is equivalent to (2) and the production noise estimate is essentially a rearrangement of (7). From these estimates, stacked vectors were formed with $L = 1, 2, 4, 6, 8, 10, 20, 30, 40$. This corresponds to assuming zero time-dependencies beyond $8, 16, 32, 48, 64, 80, 160, 240$, and 320 ms.

The MAS eigen-basis was found by computing an eigen-decomposition of the sample covariance matrix, $\hat{R}_{ML}$. The orthogonal projection matrix A was obtained by discarding eigenvectors with eigenvalues less than $V = 0.005\theta_{\max}$ where $\theta_{\max}$ is the maximum eigenvalue. The speech vectors and the production noise vectors were projected onto the MAS using A and the capacity of the speech communication channel was found by evaluating (12).

3.2. Results

Figure 1 shows the largest 100 eigenvalues of $\hat{R}_{ML}$ for $L = 40$. The plot shows that $V = 46$. This implies that the MAS can be described using a basis of 46 eigenvectors.

Figure 2 shows the result of projecting speech onto the MAS. We see that features such as pitch are suppressed, and that tempro-spectro modulations attributed to the shape of the vocal-tract are preserved. This is consistent with the notion that pitch contains no lexical information [22]. In general, the MAS projection acts as a low-pass filter that smooths the spectra. This behavior is consistent with [23] where it was found that intelligible speech could be synthesized from bandpass-filtered log-spectra.

Figure 3 shows the capacity of the speech communication channel as a function of L. We see that the capacity is approximately 100 b/s, which is comparable to the lexical information rate of speech of 50 b/s when phonemes are considered. We find that the time-dependencies are negligible for $L > 10$. This means that time-dependencies tend to last 80 ms, which is consistent with the average duration of a phoneme (e.g., [24]).

4. DISCUSSION & CONCLUSION

We developed an information theoretical model of speech communication without assuming any knowledge of speech production, language, or auditory processing. A compressed representation of speech that retained relevant linguistic information was determined by applying the information bottleneck principle to a chorus of

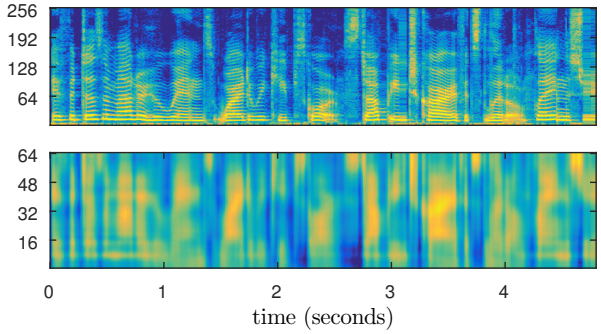

Fig. 2. Top: a sequence of log-spectra produced by a talker. Bottom: the same signal projected onto the message articulation space. The bandwidth of the signals is 8 kHz.

speech signals. From a small set of candidate representations we selected the representation that minimized the information bottleneck. The selected representation was inspired by the human auditory system. Based on this representation, the capacity of the speech communication channel was measured.

The capacity of 100 b/s of the speech communication channel that we found is approximately a factor of two larger than speech information rate estimates based on linguistic models. We believe that this factor of two can be attributed to assigning non-zero probabilities to all linear combinations of eigenvectors that define the MAS. In practice, humans cannot physically produce all linear combinations of the eigenvectors. For example, if the spectrum of one English phoneme is added to the spectrum of a different English phoneme, we do not, in general, obtain the spectrum for a third English phoneme. Moreover, the capacity of the MAS includes sounds associated with different dialects, which do not add to the speech information rate. We hypothesize that replacing the multivariate Gaussian model with a Gaussian Mixture Model could resolve this issue. If such a mixture model was used, linear combinations of eigenvectors that do not occur in the chorus would be assigned zero-probability. This would reduce the entropy rate of the message estimate, causing the capacity of the speech communication channel to drop.

We decomposed the spectral representation that minimizes the bottleneck into eigenvectors that span time and frequency by stacking successive spectra. This made it possible to evaluate the capacity of speech communication. Although we did not focus on creating a basis suitable for resynthesis, we are able to reconstruct intelligible speech signals based on the smooth auditory-spectra obtained from projections onto the MAS like the one shown in Figure 2. In part this is due to the fact that the smooth functions in the logarithmic domain can display relatively sharp transitions in the linear domain.

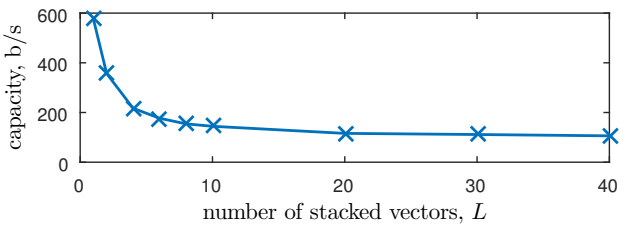

Fig. 3. The capacity of the speech communication channel as function of time-correlation duration.

5. REFERENCES

[1] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[3] W. J. Levelt, "Models of word production," *Trends in cognitive sciences*, vol. 3, no. 6, pp. 223–232, 1999.

[4] P. B. Denes, "On the statistics of spoken English," *J. Acoust. Soc. Am.*, vol. 35, no. 6, pp. 892–904, 1963.

[5] R. M. Fano, "The information theory point of view in speech communication," *J. Acoust. Soc. Am.*, vol. 22, no. 6, pp. 691–696, 1950.

[6] J. L. Flanagan, *Speech analysis synthesis and perception*, Springer, 1972.

[7] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, 2015.

[8] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, January 1947.

[9] K. D. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, November 1962.

[10] "American national standard methods for calculation of the speech intelligibility index," *ANSI/ASA S3.5-1997 (R2012)*, 2012.

[11] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, April 2005.

[12] N. Tishby, N. Pereria, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999.

[13] R. M. Hecht, E. Noor, G. Dobry, Y. Zigel, A. Bar-Hillel, and N. Tishby, "Effective model representation by information bottleneck principle," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1755–1759, 2013.

[14] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, 2006.

[15] S. Van Kuyk, W. Bastiaan Kleijn, and Richard C. Hendriks, "An intelligibility metric based on a simple model of speech communication," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[16] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep.*, vol. 35, pp. 8, 1993.

[17] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[18] W. B. Kleijn, J. B Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 43–54, 2015.

[19] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information," *Proc. IEEE. Int. Conf. Acoust.. Speech. Signal Process., (ICASSP)*, pp. 654–658, 2016.

[20] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers," in *Proc. of SPECOM*, 2006, vol. 6, pp. 431–435.

[21] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[22] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T.D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, no. 4497, pp. 947–949, 1981.

[23] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, 2009.

[24] Jan P. H. van Santen, "Quantitative modeling of segmental duration," in *Proceedings of the Workshop on Human Language Technology*, 1993, pp. 323–328.