PERCEPTUAL ENTROPY RATE ESTIMATES FOR THE PHONEMES OF AMERICAN ENGLISH

V. van de Laar, W.B. Kleijn, and E. Deprettere

Circuits & Systems Group, Department of Electrical Engineering
Delft University of Technology, Delft, The Netherlands

ABSTRACT

We estimated the perceptual entropy rate of the phonemes of American English and found that the upper limit of the perceptual entropy of voiced phonemes is approximately 1.4 bit/sample, whereas the perceptual entropy of unvoiced phonemes is approximately 0.9 bit/sample. Results indicate that a simple voiced/unvoiced classification is suboptimal when trying to minimize bit rate. We used two different methods for the entropy estimation, and the results of both methods show that short segments of unvoiced speech are approximately Gaussian.

1. INTRODUCTION

Many source-controlled multimode speech coders use phonetic classification to determine the bit allocation algorithm to be used for each segment of speech [1]. For the design of such bit allocation algorithms it is important to know the minimum bit rate needed for transparent coding of each of the phonetic classes used by the coder. This bit rate is called the perceptual entropy rate. Often, the coder uses only two or three phonetic classes, since a more detailed classification requires a long delay. Examples of phonetic classes commonly used are: voiced phonemes, unvoiced phonemes, and silence or background noise [1].

Until now, perceptual entropy rate estimates were published only on unclassified speech [2] and music [3]. Yet many source-controlled multimode coders assign more bits to voiced segments than to unvoiced segments [1, 4]. To obtain the perceptual entropy we applied Johnston's masking model [2] which is used in audio coding [5] and speech coding [6].

This paper presents the results of perceptual entropy rate estimates for the phonemes of American English. They were obtained by analyzing the TIMIT Speech Corpus [7], containing five hours of speech by native speakers of American English. Since perceptual entropy rate estimates of a single phoneme do not take into account the probability of occurrence of that phoneme, the estimates represent the conditional perceptual entropy given the phoneme.

In Section 2 we outline the methods used for perceptual entropy rate estimation. One masking model was used in conjunction with two different entropy estimation methods. Section 3 establishes the consistency of both entropy estimation methods, outlines the calculation of the entropy for unvoiced phonemes and presents our results. Finally, Section 4 presents a discussion and our conclusions.

2. PERCEPTUAL ENTROPY ESTIMATION

As a preprocessing step for perceptual entropy estimation, the entire TIMIT Speech Corpus was downsampled to 8 kHz to comply with the de-facto standard sampling rate used in speech coding. For this purpose, a 100'th order low-pass FIR filter was used with a cut-off frequency of 3.72 kHz. Then for one segment of each phoneme of every sentence in the data base, the perceptual entropy rate was estimated by a procedure consisting of two steps. First, the masking threshold was estimated using the auditory model of Johnston [2]. The masking threshold is the maximum additive noise spectrum that can be injected into the signal without being perceived. The model approximates the actual masking threshold with a staircase threshold consisting of adjacent, non-overlapping frequency bands. The bandwidth of each of the bands is equal to the critical bandwidth at the center frequency of the band. Johnston's results [2] show this staircase threshold represents a close approximation to the actual masking threshold. The masking threshold was used in the second step which consisted of the actual entropy estimation. Two different entropy estimation methods were applied: the first one is based on rate-distortion theory, and the second one uses the quantization indices of the real and imaginary part of the complex spectrum. The methods are described in the following two subsections.

2.1. Gaussianity based method

Under the assumption that the speech signal is Gaussian, the average entropy rate R of a signal with power density spectrum $\mathcal{P}(\omega)$ and an error power density spectrum $\mathcal{N}(\omega)$ equals [8]

$$R(\mathcal{N}) = \frac{1}{\pi}\int_0^{\pi} \max\left(0, \frac{1}{2}\log_2\left[\frac{\mathcal{P}(\omega)}{\mathcal{N}(\omega)}\right]\right) d\omega. \qquad (1)$$

In this case, $\mathcal{N}(\omega)$ is the masking threshold.

By accumulating estimates from segments of every occurrence of a phoneme, a histogram is generated. Such a histogram shows the distribution of the perceptual entropy of a particular phoneme, and it indicates the variability of the perceptual entropy.

The actual perceptual entropy of a speech segment is not higher than the results obtained with this method, since a Gaussian distribution is assumed and such a distribution has the highest differential entropy of all possible distributions [8]. Thus, only if speech is Gaussian does this method give the actual entropy. In case speech is not Gaussian, an upper limit to the actual entropy is obtained.

2.2. Entropy of quantization indices

The second method of entropy estimation requires quantization of the real and imaginary part of the complex spectrum in the same way as in [2]. Quantization is performed using a uniform quantizer with a frequency-dependent step size. Based on the assumption that the energy is divided equally among all spectral lines in a frequency band, and using the fact that the energy is distributed equally among the real and imaginary part of the complex spectrum, the step size is determined for each of the frequency bands of the staircase masking threshold. The quantization indices thus obtained are used to estimate the entropy according to

$$H = -\sum_{i=1}^{M} p_i \log_2 [p_i], \qquad (2)$$

where M is the number of different quantization indices and p_i is the probability that the i'th index occurs.

To account for dependence between the quantization indices of N adjacent spectral lines of the real or imaginary part of the complex spectrum, the probability of vectors of dimension N must be estimated. Then, the entropy per symbol H_N is a decreasing sequence, and its limit is the entropy H_∞ of a source with memory. When H_∞ is known, all dependence is accounted for. Since high memory requirements prohibit entropy estimation of dimensions greater than 6 or 7, the obtained curve must be extrapolated to estimate H_∞.

To estimate the probability distribution reliably, a large number of quantization indices are needed. Therefore, the quantization indices of every occurrence of a particular phoneme are accumulated. Thus, one perceptual entropy estimate is obtained for every dimension N.

Since a rectangular lattice quantizer is used, the entropy H_∞ is at least 0.255 bit/sample higher than the rate-distortion function [9]. This difference is due to the fact that this quantizer does not use an optimal packing.

In conclusion, H_∞ minus 0.255 bit/sample represents an estimate of the actual entropy of the segment. This result is lower than the result obtained from the Gaussianity based method if the speech in the segment is not Gaussian. Otherwise the two methods yield equal results.

2.3. Limitations of entropy estimation methods

The actual bit rates needed for transparent coding of "running" speech are slightly higher than the results indicate, since the nonstationary character of the signal must be taken into account. Both methods assume the signal is stationary, and to ensure the signal statistics change as little as possible within a frame, a framelength of 128 samples was used in the analysis.

The additional entropy rate required to describe the change in signal statistics is approximately 50 bits per second. This rate corresponds to the entropy of a single phoneme, which is about 5 bit/phoneme (based on the relative frequencies in [10]), and the average number of phonemes spoken per second, which is about 10. In addition, the variability of the power density spectra and masking threshold across the ensemble of one phoneme accounts for some additional entropy rate.

In the analysis the second frame of each phoneme was used, since at the onset and at the end of a phoneme signal statistics change faster than elsewhere. However, this underestimates the entropy rate because there is additional variation across the phoneme.

For voiced speech, a possible disadvantage of using short framelengths is the high mutual information between adjacent frames caused by the nearly periodic structure of the waveform. For longer frames this effect decreases, and therefore a larger framelength may be more appropriate for voiced speech.

3. RESULTS

In this section the results of both entropy estimation methods are compared. Then, the estimation of H_∞ for unvoiced phonemes is outlined, and a lower bound to the entropy of voiced phonemes is described. Finally, the last subsection presents the results for eight phonetic classes.

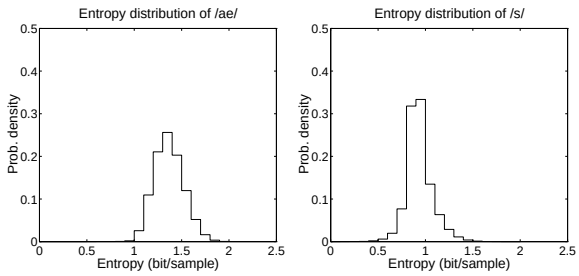

Figure 1: Histograms with the entropy distribution for the Gaussianity based method for /ae/ (left) and /s/ (right).

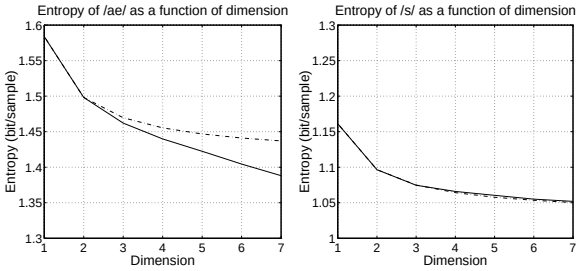

Figure 2: Entropy estimates for the entropy of quantization indices method (solid) for /ae/ (left) and /s/ (right). The dashed-dotted curves are those for a first-order Markov model.

3.1. Validation of entropy estimation methods

For a Gaussian distribution, the methods of Sections 2.1 and 2.2 must give the same outcome, since in this case Equation (1) gives the actual rate-distortion function. Unvoiced speech is likely to have a Gaussian distribution, whereas voiced speech is not.

For the phonemes /ae/ and /s/ the results of the Gaussianity based method and the entropy of quantization indices method are shown in Figures 1 and 2.

For /s/ the value of H_∞ is 1.03 bit/sample (see the following subsection) and the mean of the Gaussianity based

method is 0.89 bit/sample. If the 0.255 bit/sample offset is subtracted from H_∞ the result is slightly lower than that of the Gaussianity based method. This is consistent with the expectations at the beginning of this subsection, and it was found to be true for all unvoiced phonemes. Therefore, short segments of unvoiced speech are approximately Gaussian.

For /ae/, Figures 1 and 2 show that subtracting the offset from H_7 yields an entropy which is significantly lower than the Gaussianity based result. Again, this is consistent with the expectations at the start of this subsection. Since this was true for all voiced phonemes, the Gaussianity based results for these phonemes must be regarded as upper limits to the actual entropy. In addition, convergence of the entropy of quantization indices of /ae/ does not become apparent for dimensions smaller than eight. However, the curve for /ae/ must also converge, since H_N is a decreasing sequence which is lower bounded by zero.

To evaluate further the reliability of the estimates, we also used frames of 256 and 512 samples. Varying the frame-length did not significantly influence the estimates. Results varied by only a few hundredth bit/sample. For voiced phonemes this seems somewhat surprising, and it may be due to the limitations of the masking model which hardly takes into account the fine spectrum.

3.2. Unvoiced phonemes

H_∞ for unvoiced phonemes can be calculated by assuming a first-order Markov chain as a model of the quantization indices. This leads to a simple recursive expression that yields a curve which matches the H_N-curve very well.

For brevity, let $F_N(U)$ denote the conditional entropy of the N'th symbol u_N if the sequence of $N-1$ previous symbols $\{u_{N-1}, \ldots, u_1\}$ is known. Then the joint entropy of $\{u_N, \ldots, u_1\}$ is

$$H\left(U_N, \ldots, U_1\right) = \sum_{i=1}^{N} F_i\left(U\right), \tag{3}$$

and consequently the joint entropy per symbol is

$$H_N\left(U\right) = \frac{1}{N} \sum_{i=1}^{N} F_i\left(U\right). \tag{4}$$

From (4) it follows that

$$F_{N+1}(U) = (N+1)H_{N+1}(U) - NH_N(U) \tag{5}$$

and combined with $F_{N+1}(U) - F_N(U) \leq 0$ [11] (with equality for all $N > n$ for an n'th order Markov model) Equation (5) yields

$$H_{N+1}\left(U\right) \leq \frac{2N}{N+1} H_N\left(U\right) + \frac{N-1}{N+1} H_{N-1}\left(U\right), \tag{6}$$

with equality for $N \geq 2$ if the source is a first-order Markov chain. The curves for (6) with equality are shown in Figure 2. H_∞ is determined by substituting $N = 1$ in (5) and realizing that for a first-order Markov chain $H_\infty = F_2$. Thus,

$$H_\infty = 2H_2 - H_1. \tag{7}$$

Figure 2 shows that the quantization indices of /s/ are modeled accurately by a first-order Markov model, whereas the indices of /ae/ contain dependence such a model cannot account for. This is a general distinction between unvoiced and voiced phonemes caused by the pulse-like structure of voiced speech.

3.3. Voiced phonemes

Since the structure of voiced phonemes resembles a pulse train, the dependence between quantization indices is different than for unvoiced phonemes. A lower bound for the entropy of the quantization indices of voiced phonemes is found by substituting the original signal phase with a linear phase of random slope. The corresponding time-domain signal has a pulse train character. (The spacing of the pulses is determined by the fundamental frequency and the offset of the pulse train is determined by the slope of the phase spectrum.) By altering the phase spectrum of every segment in the ensemble and estimating the entropy with the entropy of quantization indices method, a lower bound to the actual entropy is found. The slope of each phase spectrum was a uniformly distributed random variable on $[-s, s]$, and s corresponds to the average pitch in the data base. As 70% of the sentences were spoken by male speakers, the average pitch of male and female speakers was weighted to obtain an accurate value for the slope range. The entropy curve obtained this way is shown in Figure 3.

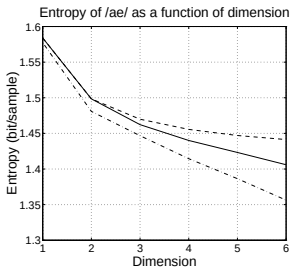

Figure 3: Entropy of quantization indices of unmodified /ae/ (solid), the first-order Markov model (dashed, top curve) and random slope linear phase (dashed-dotted, bottom curve).

Of course, the unmodified data from the data base yield a higher entropy, since for higher frequencies the power density spectrum has a more noise-like character.

The pulse train structure in the time-domain causes strong dependencies in the phase spectrum. Therefore the entropy decreases faster with increasing N than for unvoiced speech, which is nearly Gaussian. Since this decrease is not as fast as for linear phase spectra, the real phase spectrum presumably is between linear and random. In conclusion, the entropy of quantization indices for a modified ensemble with linear phase spectra with random slope gives a lower bound to the actual entropy.

3.4. Results for phonetic classes

Table 1 shows the results of the Gaussianity based method for the phonetic classes used in the manual of TIMIT data base [12]. Results for voiced phonemes must be regarded

as upper bounds to the actual entropy. This suggests that the perceptual entropy of voiced stops and (af)fricatives is lower than that of their unvoiced counterparts.

Table 1: Gaussianity based results for eight phonetic classes. A sampling rate of 8 kHz was assumed.

Phonetic class	Voicing	Entr. (b/spl)	Rate (kb/s)
Stops	+	0.96	7.68
	-	0.95	7.60
Affricatives	+ & -	1.00	8.00
Fricatives	+	0.94	7.52
	-	0.91	7.28
Nasal consonants	+	1.20	9.60
Semivow. & glides	+	1.43	11.44
Vowels	+	1.39	11.12

Table 1 also indicates that a simple voiced/unvoiced classification is suboptimal when trying to minimize bit rate. For instance, a voiced stop requires at most approximately 0.96 bit/sample, while a voiced/unvoiced classification would assign more bits to such a stop. However, a more detailed classification requires a larger delay, which is undesirable in real-time coding. This limits the use of a more detailed classification to off-line applications.

4. DISCUSSION AND CONCLUSIONS

Results for unvoiced phonemes presented in this paper differ considerably from the results of Kubin et al. [13] which indicate there is almost no perceptual entropy in unvoiced phonemes. They used a 10'th order linear predictor, and for unvoiced phonemes replaced the residual with white Gaussian noise. The residual of voiced phonemes was not altered, and the prediction coefficients were not quantized. It was found that the quality of the reconstructed speech was very good (4.0 on MOS scale). This difference for unvoiced phonemes is explained by a number of reasons. First, in [13] an MOS test was used, so that the original signal and the coded signal may sound different, although the coded signal still sounds very good. We used a masking model which only describes whether additive noise can or cannot be perceived, but in general it does not describe whether two signals are perceptually equivalent [14]. To achieve the latter more detailed auditory models must be used. In [13] no masking model was used, so that the original and reconstructed speech may differ by more than only the presence of a masked additive noise signal. Hence the results are not directly comparable. In addition, Johnston's masking model may provide a lower bound on the masking threshold, i.e., masking may be stronger than Johnston's model indicates. Finally, in [13] "running" speech coding was used, so that temporal masking, in particular postmasking [15], may play a role by rendering unvoiced phonemes less well audible.

From the results of this paper and cited papers, a number of conclusions can be drawn. First, short segments of unvoiced speech are modeled accurately by a Gaussian distribution. Second, the quantization indices of unvoiced phonemes obtained by using Johnston's masking model and quantization procedure [2] are described by a first-order Markov model. Third, the perceptual entropy of unvoiced phonemes is lower than the upper bound on the perceptual entropy of most voiced phonemes. This justifies the assignment of more bits to voiced segments than to unvoiced ones. Fourth, when trying to minimize bit rate, results shown in Table 1 indicate that a simple voiced/unvoiced classification is suboptimal. Multimode coders using a more detailed classification could exploit the differences in perceptual entropy between the phonetic classes from Table 1. Finally, results from [13] compared to those from this paper indicate that waveform-approximating coding of unvoiced speech using Johnston's masking model is unnecessarily restrictive for good perceptual quality of reconstructed speech.

REFERENCES

[1] A. Das, E. Paksoy, and A. Gersho, "Multimode and variable-rate coding of speech," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 257–288, Amsterdam: Elsevier Science Publisher, 1995.

[2] J. Johnston, "Perceptual entropy estimation using noise masking criteria," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (New York), pp. 2524–2527, 1988.

[3] R. Veldhuis, "Bit rates in audio source coding," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 1, pp. 86–96, 1992.

[4] W. Kleijn and K. Paliwal, "An introduction to speech coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 1–47, Amsterdam: Elsevier Science Publishers, 1995.

[5] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[6] J.-H. Chen and D. Wang, "Transform predictive coding of wideband speech signals," *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, pp. 275–278, 1996.

[7] "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus." NIST Speech Disc 1-1.1, National Institute of Standards and Technology, Gaithersburg, MD, October 1990.

[8] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, New Jersey: Prentice Hall, 1971.

[9] A. Gersho and R. Gray, *Vector Quantization and Signal Compression.* Dordrecht: Kluwer Academic Publishers, 1992.

[10] H. Edwards, *Applied Phonetics: the sounds of American English.* San Diego, CA: Singular Publishing Group, 1992.

[11] D. Boekee and J. van der Lubbe, *Informatietheorie.* Delft: Delftse Uitgeversmaatschappij, 1991. In Dutch.

[12] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation.* National Institute of Standards and Technology, Gaithersburg, MD 20899, February 1993.

[13] G. Kubin, B. Atal, and W. Kleijn, "Performance of noise excitation for unvoiced speech," in *Proc. IEEE Workshop on Speech Coding for Telecomm.*, (Saint-Adele, Quebec), pp. 35–36, 1993.

[14] R. Veldhuis and A. Kohlrausch, "Waveform coding and auditory masking," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 397–431, Amsterdam: Elsevier Science Publishers, 1995.

[15] E. Zwicker and H. Fastl, *Psychoacoustics, facts and models.* Berlin: Springer Verlag, 1990.