6th International Conference on Applied Informatics
Eger, Hungary, January 27–31, 2004.

Algorithm for segmenting speech sample sequences using the Jensen-Shannon divergence

István Pintér

Department of Automation and Applied Informatics,
Kecskemét College
e-mail: pinter.istvan@gamf.kefo.hu

Abstract

Recently possible applications of the entropy and different entropy-based divergence measures are investigated in many areas of information technology. In case of digital speech processing publications can be found for entropy-based segmentation of noisy speech both in time domain and in frequency domain. The application of the Jensen-Shannon divergence has also been reported in a speaker recognition problem. In this paper an algorithm is proposed using the Jensen-Shannon divergence in time-domain for segmenting noisy speech.

Categories and Subject Descriptors: C.3 [Special-purpose and application-based systems (J.7)]: Signal processing systems

Key Words and Phrases: speech segmentation, robust algorithm, Jensen-Shannon divergence

1. Introduction

The entropy-based approach has increasing importance in development of robust algorithms for nonstationary signal processing applications. From conceptual point of view there are algorithms using directly the notion of Shannon's or Rényi's entropy, but the applications of some entropy-based divergence measures have also been published. Among these there are papers on successful applications of the Jensen-Shannon divergence, the generalized Jensen-Shannon divergence or generalized Jensen-Rényi divergence. Several examples are the following: analysis of symbolic sequences [1] [2], digital image segmentaion and registration [3] [4], digital speech processing applications [5] [6] [7].
Considering the latter case in this paper, there are many efforts wordwide to develop

speech information systems using speech input. However, the incoming speech signal is usually degraded by noise and different channel distortions – therefore development of algorithms with some "immunity" contrary to above mentioned degrading effects is very important, so it is the subject of intensive research recently. In the next paragraph a succint overview of these robust, entropy-based speech processing algorithms is given as illustration.
Robust endpoint detection is important in some applications, e.g. in isolated word recognition. The algorithm, proposed in [5] uses the notion of Shannon-entropy. Following the frame-by-frame processing method – which is widely accepted in digital speech processing [8] [9] – the authors computed the entropy contour using the so called normalized and wighted spectral energy vector, and it has been used for determining the endpoints of the uttered word. As it has been reported, considering both the speech-detection accuracy and the speech recognition accuracy, the entropy-based algorithm gave better results, than the conventional energy contour based one. For a similar purpose there is another example [7], but the algorithm works directly in the time domain instead of frequency domain, by using the amplitude histogram for computing the Shannon-entropy. In the task of isolated word recognition the novel algorithm achieved better word recognition accuracy in comparison with the energy-based method. Finally, the authors of [6] proposed an algorithm for the robust speaker recognition problem, by marking the most important speech segments in verification the talker using the Jensen-Shannon divergence.
In this paper the application of the Jensen-Shannon divergence for speech segmentation is investigated.

2. The entropy and the Jensen-Shannon divergence

The basis of the algorithms mentioned in the previous section is the concept of the so called Shannon-entropy, defined on a discrete, finite probability distribution [10]. Following this notion, let $P = \{p_j,\ j = 1, ..., K\}$ be a discrete, finite probability distribution, where $K \in N$, $K \geq 1$, $0 \leq p_j \leq 1$, and $\sum_{j=1}^{K} p_j = 1$. The Shannon-entropy of the distribution is defined as $H(P) = -\sum_{j=1}^{K} p_j \cdot \log(p_j)$. Although it measures the uncertainty in bit/symbols in case of base 2 logarithm, the entropy could also be considered as a numerical value assigned for the distribution and the inequality $0 \leq H(P) \leq \log(K)$ holds for any discrete, finite probability distributions. The largest value of the Shanon-entropy could be achieved in case of uniform distribution $P = \left\{p_j = \frac{1}{K},\ j = 1, ..., K\right\}$, while it has the smallest value when the distribution is degenerate, that is there exists some $p_j = 1$, and the other values are 0. The properties of the Shannon-entropy are discussed in detail in [11] [12]. There are useful "computing rules" in the latter textbook (which could be derived from the continuous extension of the functions not defined in the given points),

in case of $r \geq 0$, $s > 0 : 0 \cdot \log \frac{0}{r} = 0 \cdot \log \frac{r}{0} = 0$, $s \cdot \log \frac{s}{0} = +\infty$, $s \cdot \log \frac{0}{s} = -\infty$. Using the entropy-concept above, the relative entropy could be defined between two probability distribution as follows. Let's denote by $P = \{p_j,\ j = 1, ..., K\}$ and $Q = \{q_j,\ j = 1, ..., K\}$ two discrete, finite, probability distributions, where $0 \leq p_j \leq 1$, $0 \leq q_j \leq 1$, $\sum_{j=1}^{K} p_j = 1$, $\sum_{j=1}^{K} q_j = 1$. The relative entropy, or the so called Kullback-Liebler divergence is defined as: $D_{KL}(P \| Q) = \sum_{j=1}^{K} p_j \cdot \log\left(\frac{p_j}{q_j}\right)$. It has the property of non-symmetry, $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$, and it has no upper bound in case of a specific degenerate distribution, which properties are not useful in many applications. However, it is possible to define the average distribution A_{PQ} of the probability distributions P and Q as: $A_{PQ} = \{a_j, j = 1, ..., K\}$, $a_j = \frac{p_j + q_j}{2}$. Then, the Jensen-Shannon divergence $D_{JS}(P, Q)$ of the distributions P and Q could be defined using the notion of Kullback-Liebler divergence as: $D_{JS}(P, Q) = \frac{D_{KL}(P \| A_{PQ}) + D_{KL}(Q \| A_{PQ})}{2}$. The Jensen-Shannon divergence could also be defined using diretly the notion of Shannon-entropy as: $D_{JS}(P, Q) = H(A_{PQ}) - \frac{H(P) + H(Q)}{2}$.

The main properties of the Jensen-Shannon divergence are given below:

(I) $0 \leq D_{JS}(P, Q)$

(II) $D_{JS}(P, Q) = D_{JS}(Q, P)$

(III) $D_{JS}(P, Q) = 0 \Leftrightarrow P = Q$

Comments:

1. It could be proven, that the inequality $D_{JS}(P, Q) \leq 1$ also holds.

2. The triangular-inequality $D_{JS}(P, Q) + D_{JS}(Q, R) \geq D_{JS}(P, R)$ does not hold for any P, Q, R discrete probability distribution triplets, therefore the Jensen-Shannon divergence could not be considered as a distance between two discrete probability distributions. However, as it is provable, for the expression of $d(P, Q) = \sqrt{D_{JS}(P, Q)}$, the triangular-inequality holds too [13].

A useful generalisation of Jensen-Shannon divergence for N discrete, finite probability distribution could be given as: $D_{JS}(P_i, \omega_i) = \sum_{i=1}^{N} \omega_i \cdot D_{KL}(P_i \| A_{P_1 \ldots P_N}) = H(A_{P_1 \ldots P_N}) - \sum_{i=1}^{N} \omega_i \cdot H(P_i)$, where the average distribution is defined as $A_{P_1 \ldots P_2} = \sum_{i=1}^{N} \omega_i \cdot P_i$, $0 \leq \omega_i \leq 1$ and $\sum_{i=1}^{N} \omega_i = 1$. One benefit of the definiton above is that the weights could be considered as free parameters. By applying the definition of the generalised Jensen-Shannon divergence for two probability distributions, the expression below could easily be written: $D_{JS}(P, Q, \omega_1, \omega_2) = H(\omega_1 \cdot P + \omega_2 \cdot Q) - [\omega_1 \cdot H(P) + \omega_2 \cdot H(Q)]$, $0 \leq \omega_1, \omega_2 \leq 1$, $\omega_1 + \omega_2 = 1$.

It is worth-while mentioning, that the generalised Jensen-Shannon divergence could be generalised further, by using Rényi' entropy $H_\alpha(P) = \frac{1}{1-\alpha} \cdot \log\left(\sum_{j=1}^{K} p_j^\alpha\right)$,

$\alpha > 0,, \alpha \neq 1$, instead of Shannon's one.

3. Speech segmentation using the Jensen-Shannon divergence

In case of speeech processing, there are many possible speech representations [8] [9] for determining the discrete, finite probability distributions necessary for computing the Jensen-Shannon divergence discussed in the previous section. The algorithm proposed in this paper uses the sampled and quantized speech signal as input; while the speech-duration data, necessary for stopping the recursive procedure, can be found in [14].

Let's denote by $A = \{a_1, a_2, ..., a_K\}$ the finite set of symbols (alphabet), and let $P = \{p_j, j = 1, ..., K\}$ be a discrete, finite probability distribution. Let's denote by S a finite sequence of symbols in A, so that $\Pr ob\{symbol a_j occurs in the sequence S\} = p_j$. In case of a finite sequence of length N, and for a suitable large N, the probability p_j could be estimated by the relative frequency as $f_j = \frac{N_j}{N} \approx p_j$, where N_j denotes the number of occurance of symbol a_j in the symbol-sequemce S. (For numerical experiments these type of sequences could be generated by using uniformly distributed pseudo-random numbers [15], thus simulating the event occuring with probability p_j, and therefore choosing symbol a_j from A as actual symbol of the symbol sequence.)

Supposing that the first n symbols (left sequence) of the finite symbol sequence of length N generated by the probability distribution $P = \{p_j, j = 1, ..., K\}$, and the remaining $N - n$ symbols (right sequence) are generated by the probability distribution $Q = \{q_j, j = 1, ..., K\}$, we can say, that there is a change point at n in the finite sysmbol sequence S. Two important questions are the following: a) in case of unknown probability distributions P and Q, and a given symbol sequence of length N, is there a change point in the sequence?, and b) if there is a change point, where is it? These questions have been investigated in detail in [1], and an algorithm, given succintly below, has also been proposed in the article for determining the change point.

Obviously, a symbol sequence of length N could be divided in two parts at $N - 1$ positions. Let's denote by $n = 1, 2, ..., N - 1$ the sequence of possible change point-candidates. Using this notation, the symbol sequence could be divided in two parts (left-sequence and right-sequence), consisting of n or $N - n$ symbols, respectively. Let's denote $P_n(S)$ and $Q_{N-n}(S)$ the relative frequencies of symbols f_j, $j = 1, ..., K$, which could be estimated using the left-sequence and the right-sequence respectively, let's also introduce the weights as $\omega_1 = \frac{n}{N}$, $\omega_2 = \frac{N-n}{N}$, and finally let's compute the sequence of the generalised Jensen-Shannon divergences $D_{JS}(n) = D_{JS}(P_n(S), Q_{N-n}(S))$ using all change point-candidates! The decision rule for the change point proposed in [1] uses the largest value of such a sequence: $D_{JS}^{\max} = \max_n \{D_{JS}(n)\}$, and for the change point n_0: $D_{JS}(n_0) = D_{JS}^{\max}$ holds.

In this paper some results are given for segmenting speech sample sequences by applying a suitable modified version of the algorithm above. In our case, the symbols are N-bit codewords, generated by sampling (using the Shannon-Kotyelnyikov sampling rule), uniformly quantizing and binary coding the speech signal. Thus the number of symbols in the alphabet is $K = 2^N$. The speech samples could thus be considered as a symbol sequence from that finite alphabet. The discrete, finite probability distributions, necessary for computing the generalised Jensen-Shannon divergence could therefore be estimated by using the amplitude-histogram of time-domain speech samples. The proposed algorithm is given below:
Input data: N bit integers (sampled, quantized and coded speech signal)
Alphabet: coded quantization levels
Distribution: amplitude, estimated with amplitude-histogram
Algorithm (batch mode):
do recursively
split speech sample sequence at n_0 for LEFT and RIGHT part
until suitable phoneme-duration constraints apply
As illustrations of the proposed method the generalised Jensen-Shannon divergence contour has been computed for clean speech and noisy speech, uttered by a female and a male speaker respectively. By examining the figures below, the segmentation property of the algorithm is obviously apparent. On Figure 1. the spectrogram of the Hungarian word [2:S], uttered by a female speaker, can be seen. The silence at the beginning of the utterance, the very rich formant structure of the Hungarian vowel [2:], the noise-like features of the fricative [S], and also the silence at the end of the uttered word could be clearly marked using the spectrogram. Figure 2. illustrates, that the local maxima of the Jensen-Shannon divergence contour could also be used to mark the above mentioned sections of the utterance in question.

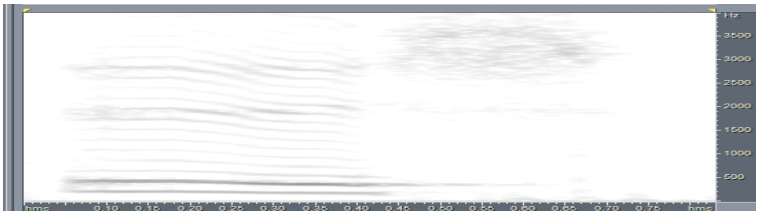

Figure 1: Spectrogram of the Hungarian word [2:S], female speaker (sampled and edited using Adobe Audition 1.0)

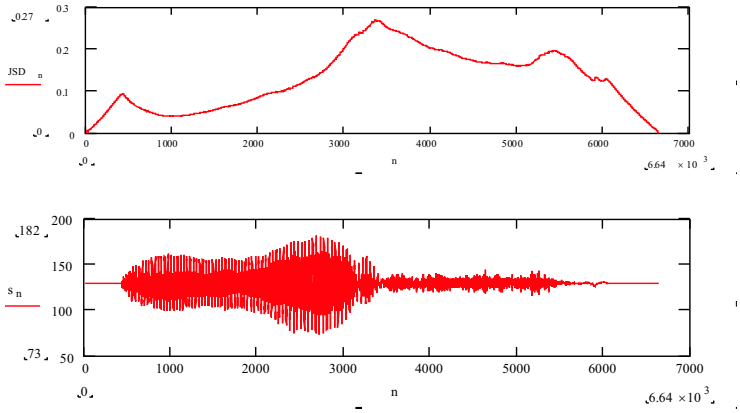

Figure 2: Hungarian word [2:S], female speaker (upper: JSD for whole utterance, lower: uttered speech (8 kHz sampling rate, 8 bit linear PCM quantization))

On the Figure 3. the spectrogram of the utterance of the English word [wVn], distorted by airplane noise, can be seen. Though it is not so easy, the endpoints of the uttered word could be denoted using the spectral representation. However, by considering the Figure 4., the local maxima of the Jensen-Shannon divergence contour could also be used to separate the noise parts in the utterance. Figure 5. illustrates the fact, that by applying the proposed algorithm, using the maximum value of the Jensen-Shannon divergence contour, the starting phase of the noisy utterance could also be detected.

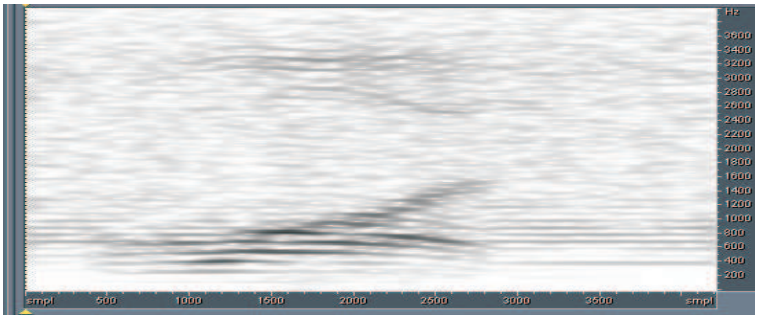

Figure 3: Spectrogram of the English word [wVn], male speaker, airplane noise (sampled and edited using Adobe Audition 1.0)

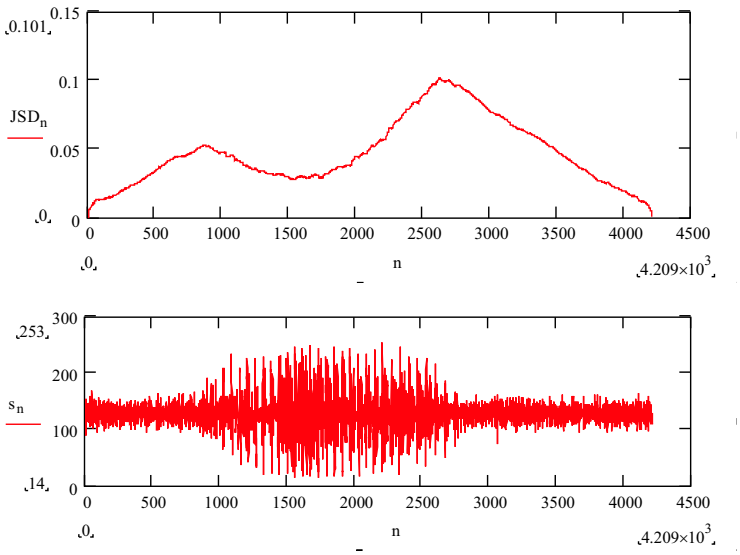

Figure 4: English word [wVn], male speaker, airplane noise (upper: JSD for whole utterance, lower: uttered speech (8 kHz sampling rate, 8 bit linear PCM quantization))

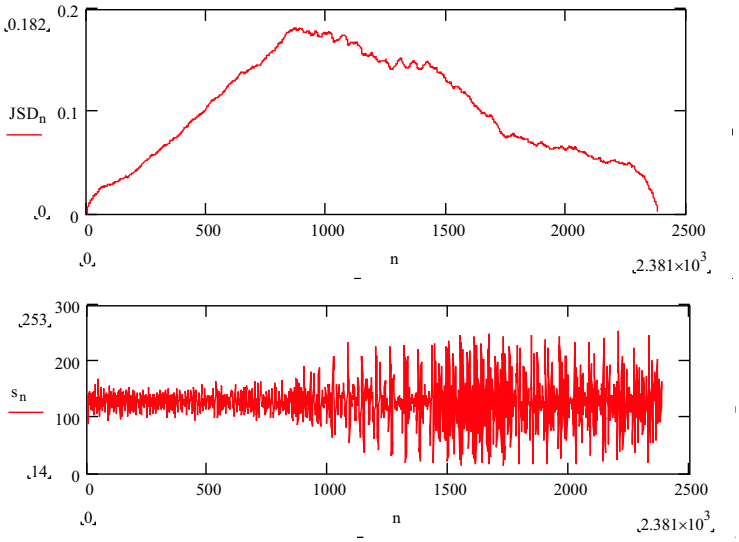

Figure 5: English word [wVn], male speaker, airplane noise (upper: JSD for left part of the utterance, lower: uttered speech (8 kHz sampling rate, 8 bit linear PCM quantization))

4. Discussion

In this paper an algorithm has been proposed for segmenting speech sample sequences using the Jensen-Shannon divergence contour, and a time-domain speech representation. The segmentation properties of the method have also been illustrated in detail.

The next step in further research is to elaborate the suitable stopping criteria using the speech duration data, available in [14]. It is also important from point of view of practical applications to develop the frame-by-frame processing version of the algorithm, using the batch mode version, proposed in this article.

References

[1] Grosse, I., Bernaola-Galván, P., Carpena P., Román-Holdán R., Oliver J., Stanley E.: Analysis of symbolic sequences using Jensen-Shannon divergence. Physical Review E, Vol. 65, pp. 041905-1-16, 2002.

[2] Korhonen, A., Krymolowski, Y.: On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. Proceedings of the 6th Conference on Natural Language Learning, Taipei, Taiwan, 2002, pp. 91-97.

[3] Gómez-Lopera, J. F., Robles-Pérez A., M., Román-Roldán R.: An Analysis of Edge Detection by Using the Jensen-Shannon Divergence. Journal of Mathematical Imaging and Vision, vol. 13. pp. 35-56., 2000.

[4] He, Y., Hamza A., B., Krim, H.: A Generalized Divergence Measure for Robust Image Registration. IEEE Transactions on Signal Processing, Vol. 51. No. 5. May 2003, pp. 1211-1220.

[5] Shen, J., Hung J., Lee, L.: Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. Procedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998.

[6] Vergin, R., O'Shaughnessy, D.: On the use of some divergence measures in speaker recognition. Proceedings of ICASSP'99, Phoenix, Arizona, 1999, Vol. 1. pp. 309-312.

[7] Waheed, K., Weaver, K., Salam, F., M.: A robust algorithm for detecting speech segments using an entropic contrast. 45th IEEE International Midwest Symposium on Circuits and Systems, Tulsa, Oklahoma, 2002, Vol. 3., pp. 328-331.

[8] Gordos G., Takács Gy.: Digital Speech Processing (in Hungarian). Műszaki Könyvkiadó, 1983.

[9] Rabiner, L., Huang, B-H.: Fundamentals of Speech Recognition. Prentice Hall, 1993.

[10] Shannon, C. E.: A mathematical theory of communication. BSTJ, Vol. 27, pp. 379-423., 1948.

[11] Csiszár, I., Körner, J.: Information Theory - Coding Theorems for Discrete Memoryless Systems. Akadémiai Kiadó, Budapest, 1981.

[12] Györfi, L., Győri S., Vajda I.: Information and Coding Theory (in Hungarian). TypoTEX Kiadó, Budapest, 2002.

[13] Fuglede, B., Topsoe, F.: Jensen-Shannon divergence and Hilbert space embedding. Preprint. University of Copenhagen, Department of Mathematics, 2003.

[14] Olaszy G.: The structure and synthesis of the most frequent elements of Hungarian speech (in Hungarian). Nyelvtudományi Értekezések, 121. sz., Akadémiai Kiadó, Budapest, 1985.

[15] Deák, I.: Random Number Generators and Simulation. Akadémiai Kiadó, Budapest, 1990.

Postal address

István Pintér
Department of Automation
and Applied Informatics
Kecskemét College
10. Izsáki str. , Kecskemét, H-6000
Hungary