The Centroid of the Symmetrical Kullback–Leibler Distance

Raymond Veldhuis

Abstract—This paper discusses the computation of the centroid induced by the symmetrical Kullback–Leibler distance. It is shown that it is the unique zeroing argument of a function which only depends on the arithmetic and the normalized geometric mean of the cluster. An efficient algorithm for its computation is presented.

Index Terms—Centroid, clustering, Kullback–Leibler distance.

I. INTRODUCTION

THE Kullback–Leibler (KL) distance originates from statistics where it is used to quantify differences between two probability distributions or densities, [5]. In information theory, it is also known as the divergence, discrimination or relative entropy. In general, the KL distance can be used to quantify differences in shape of strictly positive sequences (or functions) of which the sum (or integral) equals one. Recently, it has been used in concatenative speech synthesis to quantify the differences between spectral envelopes at concatenation points [1]–[4], [7].

The original KL distance is asymmetrical: interchanging the arguments affects the outcome. In some applications, e.g., [3], [4], [7], a symmetrical version, the symmetrical KL (SKL) distance is used.

When sequences (or functions) are clustered, e.g., in vector quantization as part of the code-book design [6] or in speech synthesis as part of the construction of a corpus [3], the centroid often represents the cluster. The centroid is the sequence (or function) with minimum average distance to the cluster. The KL centroid, induced by the KL distance, is either the arithmetic or the normalized geometric mean of the cluster, depending on which argument is subject to the minimization. Such a closed form does not exist for the SKL centroid. This paper discusses properties of the KL and SKL centroids and shows that the SKL centroid is the unique zeroing argument of a function involving only the cluster's arithmetic and normalized geometric mean. An efficient algorithm to compute it is presented.

II. DEFINITIONS AND KL PROPERTIES

The KL and SKL distances can be defined for sequences $\{p_k\}_{k\in K}$, $K \subset \mathbb{Z}$, or for functions on a continuous subset of $\mathbb{R}$. Here, sequences are considered, but similar results can be derived for functions.

For $K \subset \mathbb{Z}$, $\mathbb{P}(K) := \{p \,|\, p = \{p_k\}_{k\in K}, p_k \in \mathbb{R}, p_k > 0, \sum_{k\in K} p_k = 1\}$ denotes the set of strictly positive, sum-normalized real-valued sequences on index set K. The KL distance between $p \in \mathbb{P}(K)$ and $q \in \mathbb{P}(K)$ is defined as

$$d_{\mathrm{KL}}(p, q) := \sum_{k\in K} p_k \log \frac{p_k}{q_k}. \tag{1}$$

It is asymmetrical, i.e., $d_{\mathrm{KL}}(p, q) \neq d_{\mathrm{KL}}(q, p)$. This can be repaired by using the SKL distance, defined as

$$d_{\mathrm{SKL}}(p, q) := \tfrac{1}{2}\left(d_{\mathrm{KL}}(p, q) + d_{\mathrm{KL}}(q, p)\right) \tag{2}$$

$$= \tfrac{1}{2} \sum_{k\in K} (p_k - q_k) \log \frac{p_k}{q_k}. \tag{3}$$

Let $Q := \{q^l \,|\, q^l \in \mathbb{P}(K), l \in L \subset \mathbb{Z}\}$ denote a cluster of strictly positive sum-normalized sequences with $|Q|$ elements. The arithmetic mean $\overline{q}$ of Q is defined by

$$\overline{q}_k := \frac{1}{|Q|} \sum_{l\in L} q_k^l, \qquad k \in K \tag{4}$$

the geometric mean $\tilde{q}$ by

$$\tilde{q}_k := \left(\prod_{l\in L} q_k^l\right)^{1/|Q|}, \qquad k \in K \tag{5}$$

and the normalized geometric mean $\check{q}$ by

$$\check{q} := \frac{\tilde{q}}{\sum_{k\in K} \tilde{q}_k}. \tag{6}$$

The average left KL distance of $p \in \mathbb{P}(K)$ to Q is

$$D_{\mathrm{KL}}(p, Q) := \frac{1}{|Q|} \sum_{l\in L} d_{\mathrm{KL}}(p, q^l) \tag{7}$$

and the average right KL distance is

$$D_{\mathrm{KL}}(Q, p) := \frac{1}{|Q|} \sum_{l\in L} d_{\mathrm{KL}}(q^l, p). \tag{8}$$

The left KL centroid $\hat{q}_{\mathrm{LKL}}$ and the right KL centroid $\hat{q}_{\mathrm{RKL}}$ of Q minimize $D_{\mathrm{KL}}(p, Q)$ and $D_{\mathrm{KL}}(Q, p)$, respectively, as functions of p. It can be shown that

$$D_{\mathrm{KL}}(p, Q) = -\log \sum_{k\in K} \tilde{q}_k + d_{\mathrm{KL}}(p, \check{q}) \tag{9}$$

and

$$D_{\mathrm{KL}}(Q, p) = \left(H(\overline{q}) - \frac{1}{|Q|} \sum_{l\in L} H(q^l)\right) + d_{\mathrm{KL}}(\overline{q}, p) \tag{10}$$

with $H(q) := -\sum_{k \in K} q_k \log q_k$, which is the entropy if q is a probability distribution. The first term of (9) is independent of p and positive. Therefore, $\hat{q}_{\mathrm{LKL}} = \check{q}$. The bracketed part of (10) is independent of p and positive because $H(q)$ is concave. Therefore, $\hat{q}_{\mathrm{RKL}} = \bar{q}$. The p-dependent terms of (9) and (10) vanish for the optimal p.

The average SKL distance of a sequence $p \in \mathbb{P}(K)$ to a cluster Q is defined as

$$D_{\mathrm{SKL}}(p, Q) := \frac{1}{|Q|} \sum_{l \in L} d_{\mathrm{SKL}}(p, q^l) \qquad (11)$$

$$= \tfrac{1}{2} \left(D_{\mathrm{KL}}(p, Q) + D_{\mathrm{KL}}(Q, p) \right). \qquad (12)$$

The SKL centroid $\hat{q}_{\mathrm{SKL}}$ minimizes $D_{\mathrm{SKL}}(p, Q)$ as a function of p. Using (9) and (10), we can write

$$D_{\mathrm{SKL}}(p, Q) = C(Q) + \tfrac{1}{2} \left(d_{\mathrm{KL}}(\bar{q}, p) + d_{\mathrm{KL}}(p, \check{q}) \right) \qquad (13)$$

with $C(Q)$ only dependent on Q. Therefore

$$\hat{q}_{\mathrm{SKL}} = \arg \min_{p \in \mathbb{P}(K)} d_{\mathrm{KL}}(\bar{q}, p) + d_{\mathrm{KL}}(p, \check{q}) \qquad (14)$$

and only depends on $\bar{q}$ and $\check{q}$. Unlike $D_{\mathrm{KL}}(p, Q)$ and $D_{\mathrm{KL}}(Q, p)$, $D_{\mathrm{SKL}}(p, Q)$ has a p-dependent term that does not vanish for the optimal p. As a first guess, one could assume that $\hat{q}_{\mathrm{SKL}}$ is a linear combination of $\bar{q}$ and $\check{q}$, but a counter example is given in Section IV. Because $d_{\mathrm{KL}}(p, q)$ is convex in both arguments, it follows for $0 \leq \gamma \leq 1$ that

$$d_{\mathrm{KL}}(\bar{q}, \gamma\bar{q} + (1 - \gamma)\check{q}) + d_{\mathrm{KL}}(\gamma\bar{q} + (1 - \gamma)\check{q}, \check{q}) \leq d_{\mathrm{KL}}(\bar{q}, \check{q}). \qquad (15)$$

Therefore, $D_{\mathrm{SKL}}(\bar{q}, Q) = D_{\mathrm{SKL}}(\check{q}, Q) = C(Q) + (1/2)d_{\mathrm{KL}}(\bar{q}, \check{q})$. Of course

$$D_{\mathrm{SKL}}(\hat{q}_{\mathrm{SKL}}, Q) \leq C(Q) + \tfrac{1}{2} d_{\mathrm{KL}}(\bar{q}, \check{q}). \qquad (16)$$

A sharper upper bound is obtained by minimizing the left-hand side of (15) as a function of γ. This does not lead to a closed form for γ, but it can be shown that γ converges to $1/2$ when $\bar{q}$ converges to $\check{q}$. Another upper bound sharper than (16) is $d_{\mathrm{KL}}(\bar{q}, (1/2)(\bar{q}+\check{q})) + d_{\mathrm{KL}}((1/2)(\bar{q}+\check{q}), \check{q})$. It is the author's experience that the SKL centroid is usually close to the line segment connecting $\bar{q}$ and $\check{q}$ and that $(1/2)(\bar{q}+\check{q})$ is a good approximation of $\hat{q}_{\mathrm{SKL}}$. In fact, by using $1 - 1/x \leq \log x \leq x - 1$ and (21) in Section III, it can be shown that

$$\sqrt{\bar{q}_k \check{q}_k} \leq \hat{q}_{\mathrm{SKL}, k} \leq \frac{\bar{q}_k + \check{q}_k}{2 - d_{\mathrm{KL}}(\bar{q}, \check{q})}, \qquad k \in \mathbb{P}(K) \qquad (17)$$

if $d_{\mathrm{KL}}(\bar{q}, \check{q}) < 2$. These bounds also apply for $(1/2)(\bar{q} + \check{q})$.

Further attempts to derive a closed form for $\hat{q}_{\mathrm{SKL}}$ failed. Therefore, a numerical solution is presented in the next section.

III. ALGORITHM FOR THE SKL CENTROID

Finding the SKL centroid (14) is equivalent to the (unconstrained) minimization of

$$f(p, \lambda) := d_{\mathrm{KL}}(\bar{q}, p) + d_{\mathrm{KL}}(p, \check{q}) + \lambda \left(\sum_{k \in K} p_k - 1 \right) \qquad (18)$$

as a function of p and the Lagrange multiplier λ. The minimizing p_k can be solved from

$$\frac{\partial f}{\partial p_k} = \log \frac{p_k}{\check{q}_k} + 1 - \frac{\bar{q}_k}{p_k} + \lambda = 0, \qquad k \in K \qquad (19)$$

$$\frac{\partial f}{\partial \lambda} = \sum_{k \in K} p_k - 1 = 0. \qquad (20)$$

Because $\partial^2 f / \partial p_k^2 > 0$, for $p_k > 0$, the minimum is global. Multiplying (19) by p_k and summing over k yields $\lambda = -d_{\mathrm{KL}}(p, \check{q})$. On substitution of this into (19) and on multiplication by $p_k / \check{q}_k$ we obtain

$$\frac{p_k}{\check{q}_k} \left(\log \frac{p_k}{\check{q}_k} - d_{\mathrm{KL}}(p, \check{q}) + 1 \right) - \frac{\bar{q}_k}{\check{q}_k} = 0, \qquad k \in K. \qquad (21)$$

When $d_{\mathrm{KL}}(p, \check{q})$ is known, (21) can be solved for p. Solving

$$\frac{p_k}{\check{q}_k} \left(\log \frac{p_k}{\check{q}_k} - c \right) - \frac{\bar{q}_k}{\check{q}_k} = 0, \qquad k \in K \qquad (22)$$

with arbitrary c, however, yields a p which is not sum normalized. But $\partial p_k / \partial c > 0$, for p_k satisfying (22). Therefore, $\sum_{k \in K} p_k$ increases monotonically with c and (21) can be solved by finding the c for which the solution p of (22) satisfies

$$g(c) := \sum_{k \in K} p_k - 1 = 0. \qquad (23)$$

This c is efficiently found by the Newton algorithm starting with $c^{(0)} = -1$. In each iteration an improved approximation

$$c^{(m)} = c^{(m-1)} - \frac{g\left(c^{(m-1)}\right)}{g'\left(c^{(m-1)}\right)} \qquad (24)$$

is computed, with $g'(c) := \partial g(c) / \partial c$. In order to evaluate g and g', (22) must be solved for p. This is straightforward, since (22) can be rewritten as

$$h(x_k; y_k) := x_k \log x_k - y_k = 0, \qquad k \in K \qquad (25)$$

with

$$x_k := \frac{p_k}{\check{q}_k \exp c}, \qquad k \in K \qquad (26)$$

and

$$y_k := \frac{\bar{q}_k}{\check{q}_k \exp c}, \qquad k \in K. \qquad (27)$$

Because $y_k > 0$, $k \in K$, (25) has unique solutions for the x_k, from which the p_k follow through (26). The x_k are also easily obtained by the Newton algorithm. Starting with $x_k^{(0)} = 1$, $k \in K$, in each iteration an improved approximation

$$x_k^{(n)} = x_k^{(n-1)} - \frac{h\left(x_k^{(n-1)}; y_k\right)}{h'\left(x_k^{(n-1)}; y_k\right)}, \qquad k \in K \qquad (28)$$

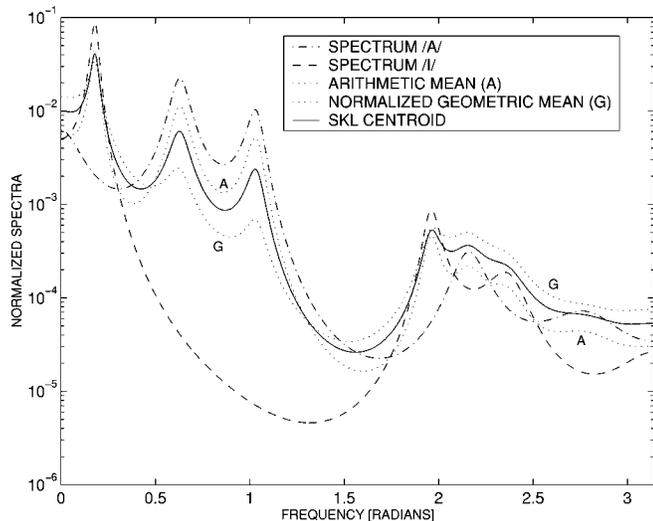is computed, with $h'(x; y) := \partial h(x; y)/\partial x$.

Fig. 1.   Spectra of vowels /a/ and /i/, the arithmetic and normalized geometric means (both plotted as dotted lines, indicated in the plot by A and G, respectively) of the spectra and their SKL centroid.

Fig. 2.   SKL centroid of vowels /a/ and /i/ (solid) together with $(1/2)(\overline{q} + \check{q})$ (dash dot) and $\hat{\gamma}\overline{q} + (1 - \hat{\gamma})\check{q}$ (dashed).

The algorithm is summarized as follows.

1) Compute $\overline{q}$ (4) and $\check{q}$ (6) from Q. Set $c = -1$.
2) Compute y_k, $k \in K$ through (27). Set $x_k = 1$, $k \in K$.
3) Compute improved estimates for the x_k, $k \in K$, by evaluating (28).
4) Continue with Step 5 if the update for x computed in Step 3 is small enough, else return to Step 3.
5) Compute p through (26).
6) Compute an improved estimate of c by evaluating (24).
7) Stop if the update for c computed in Step 6 is small enough, else return to Step 2.

IV. AN EXAMPLE

The computation of the SKL centroid of two speech spectra is presented as an example. The spectra are power-normalized all-pole spectra, estimated from Dutch vowels /a/ and /i/. The spectra, their arithmetic and normalized geometric means, and the SKL centroid computed with the algorithm from Section III are shown in Fig. 1. Fig. 2 shows a detail of $\hat{q}_{\mathrm{SKL}}$ with $(1/2)(\overline{q} + \check{q})$ and $\hat{\gamma}\overline{q} + (1 - \hat{\gamma})\check{q}$, where $\hat{\gamma}$ minimizes $|p - (\gamma\overline{q} + (1 - \gamma)\check{q})|^2$. Here, $\hat{\gamma} = 0.5260$ and $|p - (\hat{\gamma}\overline{q} + (1 - \hat{\gamma})\check{q})| = 0.0046$. Table I gives the upper bound $d_{\mathrm{KL}}(\overline{q}, \check{q})$, the improved upper bound $d_{\mathrm{KL}}(\overline{q}, (1/2)(\overline{q} + \check{q})) + d_{\mathrm{KL}}((1/2)(\overline{q} + \check{q}), \check{q})$ and the true $d_{\mathrm{KL}}(\overline{q}, \hat{q}_{\mathrm{SKL}}) + d_{\mathrm{KL}}(\hat{q}_{\mathrm{SKL}}, \check{q})$. The improved upper bound $d_{\mathrm{KL}}(\overline{q}, (1/2)(\overline{q} + \check{q})) + d_{\mathrm{KL}}((1/2)(\overline{q} + \check{q}), \check{q})$ is only 2.5% above the true $d_{\mathrm{KL}}(\overline{q}, \hat{q}_{\mathrm{SKL}}) + d_{\mathrm{KL}}(\hat{q}_{\mathrm{SKL}}, \check{q})$. This relative difference becomes even less when the term $C(Q)$ is taken into account.

Fig. 1 demonstrates the remarkable property of the normalized geometric mean and the SKL centroid that they may locally be greater than all elements of the cluster. Fig. 2 illustrates that
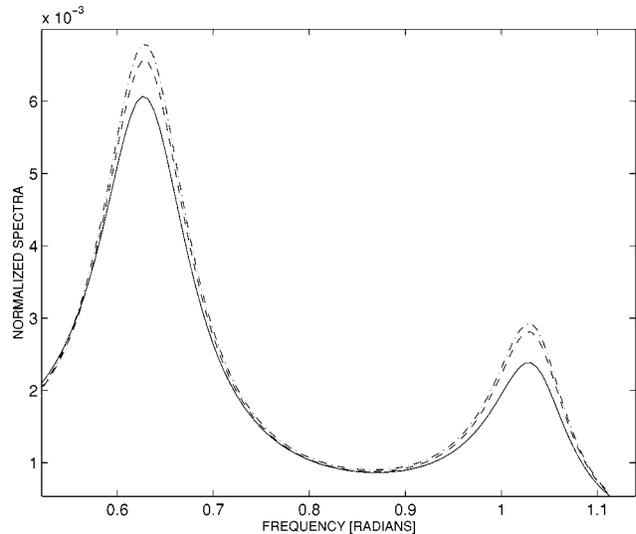
TABLE I
UPPER BOUNDS AND TRUE $d_{\mathrm{KL}}(\overline{q}, \hat{q}_{\mathrm{SKL}}) + d_{\mathrm{KL}}(\hat{q}_{\mathrm{SKL}}, \check{q})$

$d_{\mathrm{KL}}(\overline{q}, \check{q})$	0.3522
$d_{\mathrm{KL}}(\overline{q}, \frac{1}{2}(\overline{q} + \check{q})) + d_{\mathrm{KL}}(\frac{1}{2}(\overline{q} + \check{q}), \check{q})$	0.1700
$d_{\mathrm{KL}}(\overline{q}, \hat{q}_{\mathrm{SKL}}) + d_{\mathrm{KL}}(\hat{q}_{\mathrm{SKL}}, \check{q})$	0.1659

$\hat{q}_{\mathrm{SKL}}$, $(1/2)(\overline{q} + \check{q})$ and $\hat{\gamma}\overline{q} + (1 - \hat{\gamma})\check{q}$ are close but distinct. Because $\hat{q}_{\mathrm{SKL}}$ and $\hat{\gamma}\overline{q} + (1 - \hat{\gamma})\check{q}$ are distinct, the assumption that the SKL centroid is a linear combination of the arithmetic and normalized geometric mean must be rejected.

V. CONCLUSIONS

The centroid is the sequence (or function) with minimum average distance to a cluster. The KL centroid, induced by the Kullback-Leibler distance, is either the arithmetic or the normalized geometric mean of the cluster, depending on which argument is subject to the minimization. This follows elegantly on writing the average KL distance to a cluster as the sum of two positive parts, of which one depends only on the cluster and the other vanishes for the minimizing sequence. A similar expression exists for the average symmetrical Kullback–Leibler distance, but a closed expression for the SKL centroid cannot be obtained. The SKL centroid can be found as the unique zeroing argument of a function involving only the cluster's the arithmetic the normalized geometric mean. An efficient algorithm to compute it has been presented.

REFERENCES

[1] M. C. Founda, A. I. Chalamandaris, G. D. Tambouratzis, and G. V. Carayannis, "Studying the factors affecting the optimal unit selection algorithm for a TTS system for the Greek language," in *Proc. 4th Eur. Conf. Noise Control EURONOISE 2001*, Patras, Greece, 2001.
[2] M. Founda, G. Tambouratzis, A. Chalamandaris, and G. Carayannis, "Reducing spectral mismatches in concatenative speech synthesis via systematic database enrichment," in *Proc. 7th Eur. Conf. Speech Communication and Technology EUROSPEECH 2001*, vol. 2, Aalborg, Denmark, 2001, pp. 837–840.

[3] E. Klabbers and R. Veldhuis, "On the reduction of concatenation arte-facts in diphone synthesis," in *Proc. 5th Int. Conf. Spoken Language Processing (ICSLP'98)*, vol. 5, Sydney, Australia, 1998, pp. 1983–1986.

[4] ——, "Reducing audible spectral discontinuities," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 39–51, Jan. 2001.

[5] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[6] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, 1980.

[7] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. IEEE ICASSP'01*, vol. 2, Salt Lake City, UT, 2001, pp. 837–840.