TOWARDS QUANTITATIVE MEASURES OF EVALUATING SONG SEGMENTATION

Hanna Lukashevich
Fraunhofer IDMT, Ilmenau, Germany

ABSTRACT

Automatic music structure analysis or song segmentation has immediate applications in the field of music information retrieval. Among these applications is active music navigation, automatic generation of audio summaries, automatic music analysis, etc. One of the important aspects of a song segmentation task is its evaluation. Commonly, that implies comparing the automatically estimated segmentation with a ground-truth, annotated by human experts. The automatic evaluation of segmentation algorithms provides the quantitative measure that reflects how well the estimated segmentation matches the annotated ground-truth. In this paper we present a novel evaluation measure based on information-theoretic conditional entropy. The principal advantage of the proposed approach lies in the applied normalization, which enables the comparison of the automatic evaluation results, obtained for songs with a different amount of states. We discuss and compare the evaluation scores commonly used for evaluating song segmentation at present. We provide several examples illustrating the behavior of different evaluation measures and weigh the benefits of the presented metric against the others.

1 INTRODUCTION

Automatic analysis of digital audio content has become an important research field in the last years. The rapid growth of music structure analysis also poses a question of sufficient evaluation of the proposed algorithms. Commonly the automatically estimated segmentation is compared to a ground-truth, provided by human experts. Music structure annotation is a challenging task even for the human experts. Its results might strongly vary depending on a particular application or even on the cultural background of the experts. For example, for popular western music the distinguishable and repeated parts could be “intro”, “verse”, “chorus”, “bridge” and “outro”.

There are two principal approaches to the music structure analysis, namely *sequence* and *state* representation [1]. In the present work we refer only to the state representation, i.e. we consider the music audio signal as a sequence of states. Each state is characterized by a consequent time region with a similar acoustical content and is assigned to a distinct label. Thus if the similar acoustical content appears once again in the music piece it is assigned to the same label. Especially for popular western music, the semantically distinguishable and repeated parts like “chorus” or “verse” generally have constant acoustical characteristics.

Evaluating song segmentation algorithms is not a trivial task. Possible solutions for the case of state representation have been already proposed by [2], [3] and [4]. According to the knowledge of the author, up to now there is no commonly established standard way of performing a segmentation evaluation. One of the key challenges is the unknown number of possible states which might vary depending on the song. Having just one state of interest (e.g. “chorus”) [5] allows the use of *precision*, *recall* and *F-measure*, originally proposed in [6].

An ideal evaluation measure for song segmentation should possess the following properties:

- to provide the possibility to compare the results obtained by different authors andor for different algorithms;
- to be easily and intuitively understandable;
- to be insensitive to some particular properties varying for different songs, such as the number of states in the ground-truth segmentation.

It is not generally required that the estimated labels themselves should match the annotated ones, i.e. an additional stage of mapping the estimated labels into annotated ones is required. The architecture of the song segmentation system could imply that a given kind of mismatch between annotated and estimated segmentations is not considered to be a failure. For instance in [4] several annotated sequences are allowed to be mapped to a unique estimated one. Thus the evaluation measure should be able to treat this case correctly.

Representing a song as a sequence of possibly repeated states is in fact a classical clustering procedure. The same challenges appear while evaluating Speaker Clustering or Image Segmentation algorithms. As such similar evaluation measures can be used. Solomonoff et al. introduced the *purity* concept for Speaker Clustering evaluation [8], which was later extended by Ajmera et al. [9]. Another measure was originally proposed by Huang and Dom [10] for evaluating Image Segmentation. Abdallah et al. [2] adapted it to the song segmentation task. The detailed explanation and the discussion of these evaluation measures are given in the next section.

In this paper we introduce a novel evaluation measure based on the information-theoretic conditional entropy. Generally, the number and the distribution of the states is different for each song in the test set. As such the evaluation scores, obtained for different songs are not directly comparable. The proposed evaluation measure is designed to overcome this challenge. Obviously, the more states there are in the song, the more "difficult" it is to get the true segmentation by chance, or randomly. If the song has just two states of equal duration, even a "random" segmentation with two states can lead to 50% of matching between the annotated and the estimated segmentations. This percentage decreases while the number of the states is increased . The absolute value of conditional entropy itself is strongly influenced by the number and by the distribution of the states in the annotated and the estimated segmentations. Therefore an additional normalization is required. We propose using two values of normalized conditional entropies as two independent scores corresponding to the over-segmentation (errors caused by false fragmentation) and to the under-segmentation (errors caused by false merge of the segments).

2 PREVIOUS WORK

2.1 Common notations

The usual form to represent the segmentation information is to provide the starting and the ending time of each segment and to associate each segment with a particular label of the state. The same information can also be represented as a sequence of numeric labels. Estimating the automatic segmentation usually includes feature extraction from the original acoustic signal which is accompanied by a time-domain windowing. On the later stages of the segmentation algorithm additional modeling or smoothing might be done leading to the change of time discretization unit. We can apply discretization with the same time unit to the annotated ground-truth segmentation. Thus, the segmentation information can also be represented via a discrete time sequence of numeric labels. In the present work A is a sequence of labels for each unit of time discretization (for each frame) for the annotated segmentation, and likewise E is a sequence of numerical labels for the estimated segmentation.

Additionally we denote:
N - total number of frames, equal for both annotated and estimated segmentations;
N_a - number of states in the annotated segmentation;
N_e - number of states in the estimated segmentation;
n_{ij} - number of frames that simultaneously belong to the state i in the annotated segmentation and to the state j in the estimated one;
n_i^a - total number of frames, that belong to the state i in the ground-truth segmentation;
n_j^e - total number of frames belonging to the state j in the automatic segmentation.

Note, that conventionally the subscript a and the running index i are assigned to the annotated segmentation. Correspondingly, the subscript e denotes the estimated segmentation and the running index j is associated to its states.

2.2 Pairwise precision, recall, and F-measure

One of the standard metrics for clustering evaluation is pairwise precision, recall and F-measure. This technique was used for Song Segmentation evaluation by Levy and Sandler [7]. Let M_a be a set of identically labeled pairs of frames in the ground-truth segmentation, i.e. pairs of frames that belong to the same state. Likewise let M_e be a set of identically labeled frames in the estimated segmenation. Then pairwise precision (P_p), pairwise recall (R_p), and pairwise F-measure (F_p) are defined as

$$P_p = \frac{|M_e \cap M_a|}{|M_e|} \tag{1}$$

$$R_p = \frac{|M_e \cap M_a|}{|M_a|} \tag{2}$$

$$F_p = \frac{2 \cdot P_p \cdot R_p}{P_p + R_p} \tag{3}$$

where $|\cdot|$ denotes the number of the corresponding pairs.

The pairwise precision shows the accuracy of the applied segmentation algorithm due to under-segmentation, while the pairwise recall indicates the over-segmentation accuracy.

2.3 Purity concept

The *purity concept* was first proposed in [8] for the evaluation of the Speaker Clustering. Solomonoff et al. introduced a quantity measure which describes "to what extent all the utterances from the cluster came from the same speaker" [8]. The score is designed to be maximal (equal to 1) if all utterances of the cluster come from the same speaker. It reaches the lowest level of $1/k$ if the utterances are evenly distributed between k speakers. Note, that the purity concept can be easily adapted to the case of song segmentation. In our case, the word 'speaker' corresponds to the states of the annotated segmentation, and 'cluster' is assigned to the states of the estimated one.

According to the above mentioned notations the cluster purity [8] is defined as

$$r_j^e = \sum_{i=1}^{N_a} n_{ij}^2/(n_j^e)^2 \; . \tag{4}$$

Ajmera et al. [9] extended this evaluation measure using the *average cluster purity* (acp)

$$acp = \frac{1}{N}\sum_{j=1}^{N_e} r_j^e \cdot n_j^e \tag{5}$$

providing a weighted sum of single cluster purities. Furthermore they introduced the measures of *speaker purity* (r_i^a) and *average speaker purity* (asp)

$$r_i^a = \sum_{j=1}^{N_e} n_{ij}^2/(n_i^a)^2 \tag{6}$$

$$asp = \frac{1}{N}\sum_{i=1}^{N_a} r_i^a \cdot n_i^a \ . \tag{7}$$

The speaker purity estimates how well a speaker is limited to only one automatically estimated cluster. This measure is necessary to penalize a case of assigning each input utterance to a separate cluster. Ajmera et al. also suggested using a final score given as $K = \sqrt{asp \cdot acp}$.

In the case of song segmentation the above given asp measure corresponds to the over-segmentation (how well the segmentation is done due to the possible fragmentation mistakes) and likewise the acp depicts the accuracy due to the possible under-segmentation errors.

2.4 Concept of directional Hamming distance

An alternative approach to the segmentation evaluation was applied in [10] for the task of Image Segmentation. It was adapted and applied for the song segmentation task by Abdallah et al. [2].

Let T_a^i be a sequence of all frames forming a i-th state in the annotated ground-truth segmentation and T_e^j likewise a frame sequence belonging to the j-th state in the automatically estimated segmentation. The basic idea of the method is to establish the correspondence between the labels of two given segmentations by finding the sequence T_a^i with the maximum overlap for each sequence T_e^j of the estimated segmentation. The *directional Hamming distance* (d_{ae}) between annotated and estimated segmentations is given as

$$d_{ae} = \sum_{T_e^j} \sum_{T_a^k \neq T_a^i} |T_e^j \cap T_a^k| \tag{8}$$

where $|\cdot|$ denotes the duration of the overlapping parts. Normalizing the measure d_{ae} by the track length N gives a measure of the *missed boundaries*, $m = d_{ae}/N$. Accordingly the *inverse directional Hamming distance*

$$d_{ea} = \sum_{T_a^i} \sum_{T_e^l \neq T_e^j} |T_a^i \cap T_e^l| \tag{9}$$

and its normalization by the track length give a measure of the *segment fragmentation*, $f = d_{ea}/N$. The accuracy of the applied segmentation algorithm can be estimated using $1-f$ value for the under-segmentation and $1-m$ value for the over-segmentation.

We note that the evaluation measure identical to the $1-f$ score was independently proposed and applied in [4]. The author uses different mathematical expression leading to the identical numerical result. The over-segmentation mistakes were tolerated in [4], since several annotated states had been allowed to be mapped to an unique estimated one, due to an architecture of the segmentation system.

2.5 Mutual Information

In [2] Abdallah et al. proposed the information-theoretic measure (namely *mutual information*) for the segmentation evaluation. As we already mentioned in section 2.1 the segmentation information can also be represented via a discrete time sequence of numeric labels. In our notations, these are the sequences A and E for annotated and estimated segmentations correspondingly. The normalized 2D histogram of the mutual occurrence of the numeric labels in annotated and estimated segmentations can be treated as a *joint distribution* over the labels. Abdallah et al. suggested using the mutual information $I(A, E)$ as an evaluation score to measure the information in the class assignments. The mutual information is maximal when each state of the estimated segmentation maps to one and only one state of the ground-truth segmentation and it approaches zero value when the joint distribution over the labels is uniformly random. The biggest disadvantage of using the mutual information as an evaluation score is the unrestricted maximum value, which is dependent on the number of label classes and their distribution. As such the results, obtained for the songs with a different number of clusters or different distribution of the ground-truth labels, are incomparable.

3 PROPOSED METHOD

A possibility of using the conditional entropies as an evaluation score was first proposed in [2]. The authors noted that $H(A|E)$ measures the amount of ground-truth segmentation information that is *missing* in the estimated segmentation, while $H(E|A)$ measures the amount of the *spurious* information. Both $H(A|E)$ and $H(E|A)$ turn to zeros in the case of ideal segmentation. Some results of applying the conditional entropies as an evaluation score were presented in [11]. Likewise the mutual information score, discussed in section 2.5, the conditional entropies do not have a restricted maximum boundary. The higher the number of states in the segmentations (and the more uniformly these states are distributed) the higher conditional entropies.

In our method we overcome the disadvantages of using pure conditional entropy (namely, having the non-negative score with an unrestricted maximum value), and make the scores comparable for songs with a different number of clusters or different distribution of the state labels.

We denote the joint distribution (see section 2.5) of the labels (p_{ij}) and the marginal distributions for the annotated and estimated segmentations (p_i^a and p_j^e) correspondingly as

$$p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \,, \tag{10}$$

$$p_i^a = \frac{n_i^a}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \,, \tag{11}$$

and

$$p_j^e = \frac{n_j^e}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \,. \tag{12}$$

The conditional distributions are given respectively as

$$p_{ij}^{a|e} = \frac{n_{ij}}{n_j^e} \quad \text{and} \quad p_{ji}^{e|a} = \frac{n_{ij}}{n_i^a} \,.$$

Thus the conditional entropies can be written as

$$H(E|A) = -\sum_{i=1}^{N_a} p_i^a \sum_{j=1}^{N_e} p_{ji}^{e|a} \log_2 p_{ji}^{e|a} \,, \tag{13}$$

$$H(A|E) = -\sum_{j=1}^{N_e} p_j^e \sum_{i=1}^{N_a} p_{ij}^{a|e} \log_2 p_{ij}^{a|e} \,. \tag{14}$$

We propose to normalize the conditional entropies by the maximal conditional entropy for a given case. In the case of $H(E|A)$ we assume the annotated segmentation to be unchanged. The maximal conditional entropy $H(E|A)_{max}$ is achieved when the states of the automatically estimated segmentation are distributed uniformly through the states of the ground-truth segmentation. Note, that here we keep the original number of states in the estimated segmentation. In this case all the conditional distributions writes $p_{ij}^{a|e} = 1/N_e$ and the maximal conditional entropy writes

$$H(E|A)_{max} = -\sum_{i=1}^{N_a} p_i^a \sum_{j=1}^{N_e} \frac{1}{N_e} \log_2 \frac{1}{N_e} = \log_2 N_e \,.$$

Then we define the over-segmentation score (S_o) and the under-segmentation score (S_u) as

$$S_o = 1 - \frac{H(E|A)}{\log_2 N_e} \quad \text{and} \quad S_u = 1 - \frac{H(A|E)}{\log_2 N_a} \,.$$

Both scores are within the range of 0 and 1. They become maximal when the segmentations match each other perfectly, and approach a minimum value of 0 when the labels tend to be randomly chosen.

Score	Ex.1	Ex.2	Ex.3	Ex.4	Ex.5
P_p	1.00	0.53	1.00	0.53	0.55
R_p	1.00	1.00	0.23	0.62	0.55
F_p	1.00	0.70	0.38	0.57	0.55
asp	1.00	1.00	0.42	0.75	0.56
acp	1.00	0.50	1.00	0.54	0.56
K	1.00	0.71	0.65	0.64	0.56
$1-f$	1.00	1.00	0.42	0.75	0.67
$1-m$	1.00	0.58	1.00	0.58	0.67
$H(E\|A)$	0.00	0.00	1.69	0.50	0.92
$H(A\|E)$	0.00	1.09	0.00	0.94	0.92
$I(A,E)$	1.90	0.81	1.90	0.96	0.08
S_o	1.00	1.00	0.53	0.68	0.08
S_u	1.00	0.53	1.00	0.60	0.08

Table 1. Values of the different evaluation scores obtained for the examples 1-5

4 EXAMPLES

In this section we bring several examples to illustrate how the evaluation measures treat the typical errors appearing during song segmentation. For each of the examples we provide the schematic representation of the annotated and the estimated segmentations. The resulting evaluation scores for all examples are presented in Table 1.

Example 1

d	b	a	b	a	c	b	a	e	annot.
a	b	c	b	c	d	b	c	e	estim.

This example represents the ideal song segmentation for the typical song structure, consisting of the "chorus" (label a), "verse" (label b), "bridge" (label c), "intro" (d), and "outro" (e) in the annotated segmentation. In this case both errors of over-segmentation and under-segmentation are absent. As we see in the Table 1, all evaluation scores treat the case correctly. The mutual information reaches the maximal value of 1.90 bit, which is determined by the number and the distribution of the states for a given segmentation structure.

Example 2

d	b	a	b	a	c	b	a	e	annot.
a	b				a	b		a	estim.

Song segmentation algorithms are often realized by means of hierarchical agglomerative clustering. The challenging task is to terminate the agglomeration at the required level of the segmentation hierarchy, corresponding to the ground-truth reference segmentation. This example illustrates the case, when the agglomeration algorithm is terminated too late, and the resulting states are too generalized. As such there is no over-segmentation errors, but under-segmentation scores are rather poor. Note, that the mutual information score (0.81 bit) is informative only in comparison to the maximal mutual information for a given ground-truth segmentation, obtained in Example 1.

Example 3

<table>
<tr><td>d</td><td>b</td><td colspan="2">a</td><td>b</td><td colspan="2">a</td><td>c</td><td>b</td><td colspan="2">a</td><td>e</td><td>annot.</td></tr>
<tr><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td><td>f</td><td>g</td><td>h</td><td>i</td><td>j</td><td>k</td><td>l</td><td>estim.</td></tr>
</table>

The opposite situation occurs if the automatically estimated segmentation tends to assign each input utterance to a separate state. That could happen if the agglomerative clustering is terminated at a very early stage. In this case there are no under-segmentation errors, but the over-segmentation scores are low.

Note, that the over-segmentation scores (asp, $1 - f$ and S_o) and the under-segmentation scores (acp, $1 - m$ and S_u) are in good correspondence for both Example 2 and Example 3. Contrastingly, the mutual information score for the Example 3 is close to the one obtained for the "ideal" segmentation in Example 1. The mutual information $I(A, E)$ can be written as

$$I(A, E) = H(E) - H(E|A) \qquad (15)$$

where $H(E)$ is a marginal entropy of the automatically estimated segmentation. In this case a marginal entropy for 12 uniformly distributed states is really high (3.59 bit). As such even a high value of conditional entropy $H(E|A)$ (1.69 bit) cannot compensate it. The comparison of the Example 1 and Example 3 shows that a single score of mutual information is not directly applicable as an evaluation measure when the number of the states in the segmentations is changed.

Example 4

<table>
<tr><td>d</td><td>b</td><td colspan="2">a</td><td>b</td><td colspan="2">a</td><td>c</td><td>b</td><td colspan="2">a</td><td>e</td><td>annot.</td></tr>
<tr><td>a</td><td colspan="2">b</td><td>c</td><td colspan="2">b</td><td colspan="2">c</td><td colspan="2">b</td><td>c</td><td>a</td><td>estim.</td></tr>
</table>

In most cases the estimated segmentation contains both over-segmentation and under-segmentation errors. For instance it happens when the borders of the segments are not determined correctly. In the presented example the second half of the "chorus" together with the "bridge" (label c in the estimated segmentation) is recognized as a distinct state, while the first half of the "chorus" is merged with the "verse" (label b in the estimated segmentation).

Clearly, the obtained estimated segmentation in Example 4 is not ideal, but it is obviously more appropriate than the estimated segmentation in Example 3. Note, that the measure K stays nearly unchanged, and even slightly decreases.

Example 5

<table>
<tr><td colspan="3">b</td><td colspan="3">a</td><td colspan="3">b</td><td colspan="3">a</td><td>annot.</td></tr>
<tr><td>a</td><td>b</td><td>a</td><td>b</td><td>a</td><td>b</td><td>a</td><td>b</td><td>a</td><td>b</td><td>a</td><td>b</td><td>estim.</td></tr>
</table>

The structure of the ground-truth segmentation in Examples 1-4 is typical (but not obligatory!) for Rock or Pop songs. This example depicts an annotated segmentation that could often appear e.g. for Jazz or Classical music pieces. As it was noted in section 1, the evaluation score should provide a possibility to compare the segmentation efficiency for different songs. In this case the estimated segmentation tends to be random. Obviously, this estimated segmentation is not of use in practical applications. In Table 1 we see, that the over-segmentation and the under-segmentation scores (S_o and S_u) are really low while the other evaluation measures (asp, acp, $1 - f$, $1 - m$) keep relatively high values. As such the values of $1 - f$ and $1 - m$ become comparable for the Examples 4 and 5, nevertheless that the estimated segmentation in Example 5 evidently carries less information about the corresponding ground-truth segmentation.

5 DISCUSSION

Analyzing the results listed in Table 1 we can summarize the following trends. First of all, the evaluation scores discussed in section 2 are strongly dependent on the amount and the distribution of the states in both annotated and estimated segmentations. Even within one musical genre these parameters differ for each song in the test set. Evaluating a song segmentation algorithm commonly implies calculating the evaluation scores for every entry of the test set and then providing the overall results. The latter becomes impossible if the evaluation score depends on the parameters that are different for every item of the test set.

We should mention that the higher the number of the states in the segmentations the more reliable the evaluation scores asp, acp, $1 - f$ and $1 - m$. Contrastingly, if the song consists only of a few states, even a "random" segmentation yields relatively high values.

The results show the benefits of treating the over-segmentation and the under-segmentation scores independently. If one of the scores is close to 1 and the other is relatively low, then we can assume that the automatic segmentation is performed on the other level of the segmentation hierarchy. In the worst, all states of the annotated segmentation are mapped into one state of the estimated segmentation, or vice versa each frame (time discretization unit) of the song forms a distinct state in the estimated segmentation. In the latter cases the merging score K leads to the spurious results. For instance in the Examples 2 and 3 the score K indicates inconsistently high results. Therefore we believe using two independent scores S_o and S_u to be more appropriate for the case.

The comparison of the Examples 1 and 3 shows that the mutual information $I(A, E)$ is an unreliable evaluation score, especially if the number and the distribution of the states in both annotated and estimated segmentations are changed significantly.

The proposed evaluation score shows reliable results for all five presented examples. It is insensitive to changing the number of states in the segmentations and as such enables the comparison of automatic evaluation results, obtained from different songs. The scores can be loosely treated as an accuracy rate of the applied segmentation indicating the over-segmentation and under-segmentation errors. As a disadvantage of the S_o and S_u scores one can point out that in the strict sense these scores cannot be treated as the accuracy rates. The proposed normalization only restricts the boundaries of the scores and brings it within the range of 0 and 1. In point of fact the conditional entropies $H(E|A)$ and $H(A|E)$ are not a linear functions and thus a boundary restriction does not form the accuracy rate out of it. Therefore further investigations and a more complicated normalization scheme are needed.

6 CONCLUSION

In this paper we presented a novel approach to the song segmentation evaluation. The proposed evaluation score is based on the information-theoretic conditional entropies for comparing the estimated song segmentation with the one annotated by human experts (in the case of a state representation). Having applied the normalization scheme we formed the over-segmentation and the under-segmentation scores reflecting the accuracy rate of the obtained automatic segmentation given the annotated ground-truth. We compared our evaluation method to the commonly used approaches. By providing the illustrating examples we demonstrated the challenging points of the evaluation procedure and showed that the proposed over-segmentation and under-segmentation scores depicted more reliable results in comparison to the other evaluation measures.

By means of over-segmentation and under-segmentation scores one can compare the song segmentation efficiency obtained for the different songs, even when the number and the distribution of the states for these songs are various.

Since song segmentation is a particular case of classical clustering procedure, the proposed approach can be also applied for evaluating other clustering tasks like Image Segmentation or Speaker Clustering.

7 ACKNOWLEDGMENTS

This work has been partly supported by the PHAROS and the DIVAS projects, funded under the EC IST 6th Framework Program. Furthermore, the work on this publication is supported by grant No. 01QM07017 of the German THESEUS program.

8 REFERENCES

[1] Peeters, G. “Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: Sequence and State Approach”, *in Lecture Notes in Computer Science*, Springer-Verlag, 2004.

[2] Abdallah, S. Noland, K., Sandler, M., Casey, M., and Rhodes, C. “Theory and evaluation of a Bayesian music structure extractor”, *in Proc. ISMIR*, London, UK, 2005.

[3] Paulus, J. and Klapuri, A. “Music structure analysis by finding repeated parts”, *in Proc. AMCMM*, Santa Barbara, California, USA, 2006.

[4] Peeters, G. “Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum-Likelihood Approach”, *in Proc. ISMIR*, Vienna, Austria, 2007.

[5] Goto, M. “A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station”, *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[6] van Rijsbergen, C. J. *Information Retrieval*, Butterworths, London, UK, 1979.

[7] Levy, M., Sandler, M. “Structural Segmentation of musical audio by constrained clustering”, *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[8] Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H. “Clustering Speakers by Their Voices”, *in Proc. IEEE ICASSP*, Piscataway, New Jersey, 1998.

[9] Ajmera, J., Bourlard, H., Lapidot, I., and McCowan, I. “Unknown-Multiple Speaker Clustering Using HMM”, *in Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[10] Huang, Q. and Dom, B. “Quantitative methods of evaluating image segmentation”, *in Proceedings of the International Conference on Image Processing*, Washington, DC, USA, 1995.

[11] Abdallah, S., Sandler, M., Rhodes, C., and Casey, M. “Using duration models to reduce fragmentation in audio segmentation”, *Machine Learning*, 2006.