Synchronization and comparison of Lifelog audio recordings

Nielsen, Andreas Brinch; Hansen, Lars Kai

Published in:
IEEE Workshop on Machine Learning for Signal Processing, 2008.

Link to article, DOI:
10.1109/MLSP.2008.4685526

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

SYNCHRONIZATION AND COMPARISON OF LIFELOG AUDIO RECORDINGS

Andreas Brinch Nielsen, Lars Kai Hansen

Technical University of Denmark
DK-2800, Kgs. Lyngby, Denmark
{abn,lkh}@imm.dtu.dk

ABSTRACT

We investigate concurrent 'Lifelog' audio recordings to locate segments from the same environment. We compare two techniques earlier proposed for pattern recognition in extended audio recordings, namely cross-correlation and a fingerprinting technique. If successful, such alignment can be used as a preprocessing step to select and synchronize recordings before further processing. The two methods perform similarly in classification, but fingerprinting scales better with the number of recordings, while cross-correlation can offer sample resolution synchronization. We propose and investigate the benefits of combining the two. In particular we show that the combination allows sample resolution synchronization and scalability.

1. INTRODUCTION

Lifelogs are extended digital recordings of a persons life. This could for example include (e)mail correspondence, visited web sites, documents, chat logs, and video and audio recordings. The typical and original aim [1] of such recording is to boost recollection of events. Modern examples include MyLifeBits [2], LifeStreams [3] and Lifelogs [4]. Extensive digital recording could also be used for modelling behaviour as in, e.g, [5].

The task of collecting and processing Lifelog data stores is huge, and here we focus on audio aspects, as pioneered by Ellis and coworkers, see e.g., [6]. While conventional Lifelogs concern organization of personal archives we are particularly interested in the group perspective, and thus expand the scenario from including only the recordings of a single individual to integrate the recordings of multiple subjects. Multi-subjects audio analysis has been pursued earlier, e.g., in the context of conversational patterns as in [7]. We envision a setup in which employees wear microphones recording continuously while at work. Because we imagine microphones worn by individuals we can not only say something about *who said what*, but also estimate who actually received given information, i.e., *who heard what!* Here we will not be concerned with the obvious ethical issues involved in storing such audio but only investigate the mounting technical challenges.

Signal processing of multi-microphone recordings has a significant literature, see e.g., work on signal separation [8], and also includes work on distributed microphone arrays from specially equipped rooms [9]. Common to most of these is that the recordings are well synchronized and that they are recorded within the given locality thus in principle contains the same acoustic enviroment, in addition it is often assumed that the actual microphone placement is fixed and known. In our setting of 'moving microphones' some of the parameters must be inferred from the data itself. In this paper we will consider two aspects of concurrent Lifelogs, namely 1) to classify recordings as being from within the same area, meaning, that they have recorded the same audio events, and 2) we will investigate synchronization of recordings. Synchronization is necessary, because of the distributed nature of the recordings and wanted for subsequent blind signal separation processing. Recording devices that are not linked are likely to produce timing differences of the order of seconds, which will make un-mixing filters invoked by 'convolutive' blind signal separation algorithms prohibitively long.

The paper is organized as follows, in section 2 we describe two different audio similarity measures, cross-correlation and fingerprinting. In section 3 the classification problem is described. Different approaches are investigated, including one-on-one classification, a joint approach using both similarity measures and a joint classification scheme that assures a block diagonal mixing matrix. In section 4 experiments are performed within a large public data set from the AMI corpus [9] and own real-room experiments.

2. AUDIO SIMILARITY MEASURES

In this section the two measures of similarity will be presented. The normalized cross-correlation coefficient is a well known statistical quantity. The fingerprinting procedure is less so and was originally presented to identify pop songs recorded with a cell phone and compared to a large database.

1-4244-2376-7/08/$20.00 ©2008 IEEE

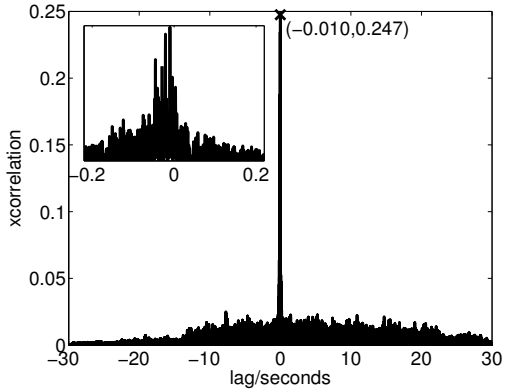

Fig. 1. The comparison of two recordings of the same meeting as a function of the lag using the normalized cross-correlation. The SNR is very good in this example, although multiple peaks exist showed in the zoomed view (insert). Only the value and location of the maximum is used as indicated.

The two methods return a measure of the similarity between two signals. A binary decision is necessary deciding whether the two signals have been recorded from the same environment. This will be achieved through the training of a classification algorithm and will be presented in the next section.

Cross-correlation. The sampled cross-correlation function is given as,

$$xc(m) = \frac{1}{N\sigma_1\sigma_2}\sum_n x_1(n)x_2(n-m),$$

where the cross-correlation is normalized with the product of the standard deviations (σ_1, σ_2) such that the range is $[-1; 1]$. As the signals have different delays and possibly have been filtered differently through different sound paths, a negative cross-correlation is as significant as a positive and thus the absolute magnitude is used. This produces a measure in the range $[0; 1]$, where the value 1 is for similar signals which surely come from the same environment while the value 0 is no correlation at all, meaning that the recordings are likely to come from different audio environments.

The cross-correlation is computed as a function of lags in the range $[-10s; 10s]$. The cross-correlation coefficients of two recordings from the same room as a function of the lag is shown in fig. 1. The lag location and value of the maximum are found. The maximum value will be used to make the binary decision, and the lag can be used to minimize the delay between recordings.

Fingerprinting. This method was proposed in [10] and used in [11]. It was originally intended for recognizing songs from short cell phone musical recordings. The fingerprint method preprocesses a recording in a way to drastically reduce the dimensionality and ease the accessibility for future comparisons.

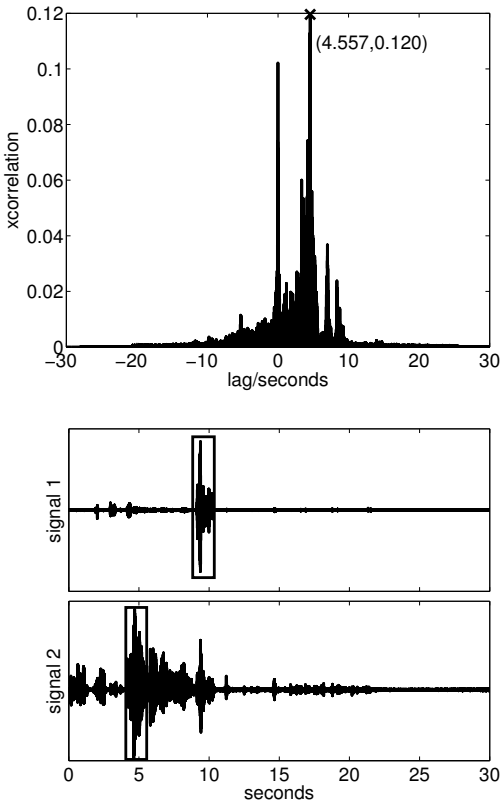

Fig. 2. The top plot is an example of the cross correlation returning an ambigous result. A peak is present in the correct place for the example (0 s lag), but another larger peak is also present (marked by an x). In the bottom plot the two signals are shown, and by investigation it was found that, by coincidence, the boxed parts are better correlated than the actually synchronized parts, thus creating an incorrect inference of the lag.

For each recording a number of *hashes* are generated, together with their associated time stamps. When two recordings are compared, all hashes are compared and when hits are found, the time difference between their time stamps is saved. A histogram is made of the time differences. If two recordings are from the same environment a relative large number of hits is expected to occur with the same time difference, and this will show up in the histogram as in fig. 3. The histogram is processed as the cross-correlation, i.e., the maximum is found and saved together with the value of the lag. While the cross-correlation function produces a lag with 'sample resolution' the resolution in the fingerprinting procedure depends on histogram and hash settings and in the current setup amounts to approximately 50 ms.

The hashes are generated from landmarks in the spectrogram. Each frequency bin is normalized (over time) to zero mean, which reduce the effect of the in general higher

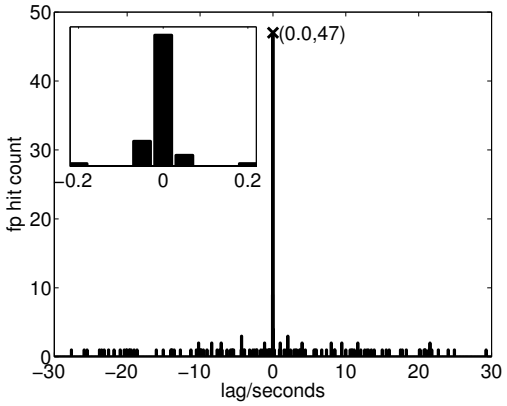

Fig. 3. The same setup as in figure 1 but using the fingerprinting method instead. Obviously the two methods return similar results, but they differ greatly in the calculations. The same clear SNR is present, but because of the lower resolution only a single peak is present in the zoomed view.

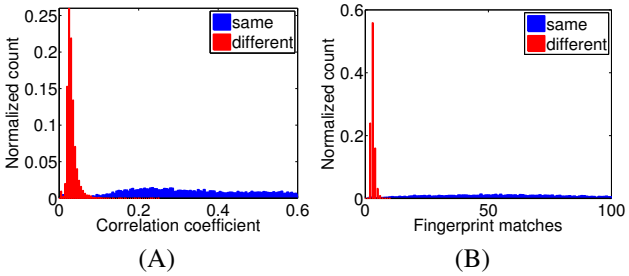

Fig. 4. A histogram of the maximum values from figure 1 over multiple windows. (A) is the cross-correlation method and (B) is the fingerprinting method. Red is for recordings from different environments and blue is for recordings from the same environment. For both methods the two classes are clearly distinguishable, and a threshold could be found by inspection of the graph.

energies in the lower frequency bands. To ensure a more uniform distribution of landmarks over the spectrogram, it is coarsely divided in both time and frequency in a number of equally sized parts. In each of these parts local maxima are found. The k largest local maxima in each part is recorded. We found that a small amount of smoothing, prior to locating maxima improved results, thus a 3x3 moving average filter was applied.

Within the parts of the spectrogram each unique pair of the k landmarks ($k(k-1)/2$ pairs) is used to generate a hash. Each hash consists of three b bit values; the (absolute) time difference between the two landmarks, the frequency value of the first landmark and the frequency value of the second landmark. All three values are discretized to b bit, and concatenated into a $3b$ bit hash. The time point of the first landmark is saved together with the hash.

3. CLASSIFICATION OF AUDIO SIMILARITY

Previously, measures of similarity between two recordings was described. The next step will be to decide when the measured similarity is significant and the recordings are considered to be from the same environment, which is a classification problem. The measures are one dimensional for both methods. Histograms of the similarities of the training set of both methods are shown in figure 4. Clearly we are looking for a threshold between the two classes. Because of the one dimensionality of the measures this can be done by simple line search minimizing the classification error rate on a training set.

To estimate the classification error rate we test the system on audio from the same and from different environments. In discriminative classification, if the individual classes do not contain the same number of samples in the training set, and this does not reflect a 'prior', it is important to normalize the classification error during training. Otherwise skewed results will be obtained. In this case it is straightforward, the number of false negatives and false positives are simply divided by the appropriate sample sizes before being added together to compute the classification error rate.

3.1. A combined approach

The two methods reviewed in the previous section differ in the resolution of the delay and in scalability with increasing number of sources. The experiments will show that the two methods perform comparably in classification and this would point to recommending cross correlation because of the increased resolution. However, when comparing recordings one-on-one, the number of comparisons will always increase quadratically with the number of recordings. This is the situation for both methods, but the complexity of the comparisons differ. The cross correlation has all the complexity in the comparison stage and is therefore severely hurt for many recordings. The fingerprinting method preprocesses the data to make the comparisons relatively light. The preprocessing is heavier than for the cross-correlation, but the preprocessing only scales linearly with the number of recordings and therefore, for increasing number of recordings the fingerprinting method will perform significantly faster. This is illustrated in fig. 5.

To use this timing advantage a combination is proposed, working in two stages. In the first stage the fingerprinting method is used to make a coarse classification. In the second stage the cross correlation is used only on the recordings that were classified as coming from the same environment in the first stage. The results from the cross correlation are used both to check the classification and the increased delay resolution is used to precision synchronize the recordings.

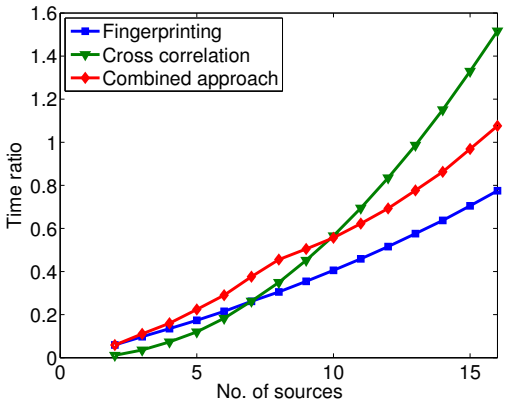

Fig. 5. The time consumption of the different proposed algorithms. Time ratio means the time it takes to process one unit of time - 1 means that processing time equals the duration of the recordings. A quadratic increase can be observed for cross correlation and a linear increase for the fingerprinting method. The proposed combination results in time consumption in between the two.

3.2. Joint classification of multiple recordings

In the previous section, classification was done based on pairwise comparison. This could result in source 1 and 2 being similar, source 2 and 3 being similar, but source 1 and 3 being dissimilar. How do we interpret this result? We need to group the sources consistently and we will not allow overlapping clusters. First the similarity measures are set up in a similarity matrix containing all the two by two comparisons. In this matrix we need to cluster the recordings resulting in a block diagonal similarity matrix (or a permutation of one).

A block diagonal form will be obtained by a greedy procedure similar to Ward's agglomerative clustering [12]. Because the similarity matrix is symmetric only the upper triangle of the matrix is considered. First, we locate the maximum similarity and the two involved recordings are grouped together. This cluster will not be split again and therefore the mean of similarities of the two recordings to the remaining recordings are calculated and entered in the similarity matrix. Then the next maximum is found, and connected either to the existing group or to another single source thereby creating a new cluster. This procedure is continued until a threshold is reached. This threshold will be trained using line search, and the same normalized classification error rate measure from before. A simple example is shown in table 1.

Common time reference. Hitherto we have concentrated on quantifying the similarity of recordings. The focus of this part will be on the timing of already classified recordings. When the recordings have been classified they can be used in other algorithms such as ICA, but many of these algorithms work better the smaller the delays between

$$\begin{bmatrix} 0 & 7 & 2 & 4 \\ 7 & 0 & 6 & 4 \\ 2 & 6 & 0 & \mathbf{8} \\ 4 & 4 & 8 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & \mathbf{7} & 3 & 3 \\ 7 & 0 & 5 & 5 \\ 3 & 5 & 0 & 0 \\ 3 & 5 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 4 & 4 & 0 & 0 \\ 4 & 4 & 0 & 0 \end{bmatrix}$$

Table 1. Block diagonal classification. 8 is the maximum and recording three and four are clustered together, and the mean of the similarities to other recordings are calculated. Next, 7 is the maximum and recording one and two are clustered. For a threshold larger than four the clustering ends here, otherwise a final step will join the two clusters.

the recordings are. For each pairwise comparison the delay is found, but these do not necessarily match each other. Therefore the first recording will be selected as the basis and the delays to the remaining recordings will be found. If more than two recordings exist there will be more lags than there are recordings. For example for four recordings there is three delays to be found while six lags have been measured. In the present experiments we use a simple least squares fit to estimate the minimal delay mismatch configuration.

4. EXPERIMENTS

The AMI corpus [9] is a large collection of multimodal meeting recordings including multiple audio recordings with different microphone configurations including microphones attached to individuals (lapel microphones). Two meetings have been used for training and two for testing. Each meeting has four participants and the recordings are from the lapel microphone, so four channels are available from each meeting. Two meetings are used together to make eight channels from two different settings.

The data is framed in 30 second frames, which are overlapping by 20 seconds. The training set was a little more than two hours long, giving 786 frames. The test set was 87 minutes long giving 524 frames. The recordings are downsampled to 8 kHz, and the range of possible delays is set to $10s$.

For the fingerprinting method the spectrogram is divided into smaller parts as explained previously. The frequency axis is divided in two and the time axis is divided into 1 s long windows (30 windows). In each part $k = 10$ landmarks are found, and a total of $k(k-1)/2 \cdot 2 \cdot 30 = 2700$ landmarks are generated per frame. The spectrogram is computed with 256 samples and 195 samples overlap, so that each part used for the landmarks becomes 64×127, and the three parts of the hash are discretized to $b = 6$ bit each.

To simulate less ideal situations, e.g., poor microphones or recording devices, uncorrelated gaussian noise was added in different signal-to-noise ratios (SNR). The SNR is calculated within each 30 s window.

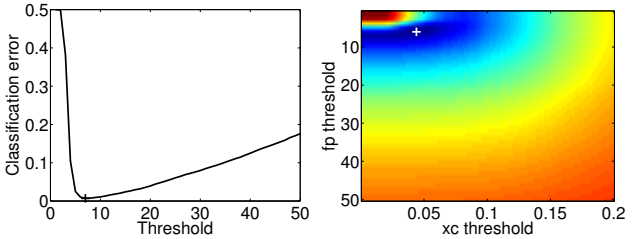

Fig. 6. (A) Line search for threshold for fingerprinting method. (B) Grid search for threshold for the combined approach. In both plots the minimum is marked with '+'. In the combined approach the cross correlation threshold is moved from just below 0.01 when only cross correlation is used to arout 0.05 in the combined case, whereas the fingerprint threshold moved from a value of 7 to 6.

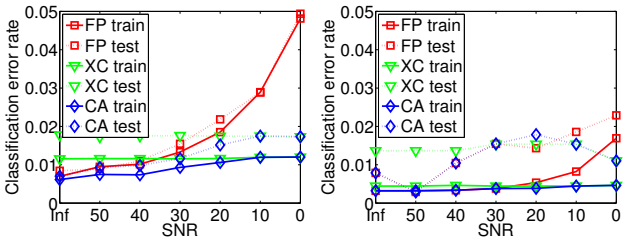

Fig. 7. Comparison of the two methods. SNR controlled by adding gaussian white noise. (A) is one by one classification. Fingerprinting takes a big hit for increasing noise, but is best in low noise. The combined approach outperforms both. (B) Forced block diagonal structure. Again the combined approach is best, except for one point (20 db).

Training was done using line search and grid search for the combined approach. In fig. 6 an example of such training can be seen. It can be seen that besides the computational advantage, an additional small reduction of the training error can be gained, since the minimum exploits the combined model space. The cross correlation threshold is driven closer to zero, whereas the fingerprint threshold only moves slightly, compared to the case when only one of the measures is used.

4.1. Results

Obviously the classification performance is important, but because of the extensive data sets resulting from Lifelogs execution time is also a concern.

Results are shown in fig. 7 (A). For low noise conditions, the fingerprinting algorithm provides the best performance of the two algorithms with a test error rate below 0.01. The method takes a significant hit in performance for additive noise and ends up around 0.05 in test error rate. Cross correlation works significantly worse with a test error rate close to 0.02 which is around twice as much as the fingerprint method. The algorithm is however quite insensitive to noise due the robustness against addition of uncorrelated gaussian noise.

If the method is used as a preprocessing step before ICA or similar evaluations, we are interested in clustering the recordings, and a block diagonal classification is relevant. The results in this setting is shown in figure 7 (B). Similar trends are found, but the fingerprinting is lot less sensitive to noise in this case. In both figures the combined approach is plotted as well. For training we expect that it performs better than the other two basic methods. We see that the performance also translates to an improved test error. This means that the good performance in low noise conditions of the fingerprinting procedure and the good performance of cross correlation in relatively high noise settings both can be achieved in the combined approach.

In figure 5 the time complexities of the algorithms are shown. The cross correlation scheme shows a quadratic growth in time, but the fingerprinting method shows close to linear growth. The reason for this is, that most of the processing time of the fingerprinting method lies in the computation of the fingerprints, whereas the actual comparisons are very fast because of the hashing structure and the severely reduced data size. In the present case, the reduced size is actually the only explanation, since a proper index structure was not used to facilitate the efficient search. The number of fingerprint computations only increases linearly with the number of sources, hence the linear increase in computational time. The combined approach uses this fact, since the fingerprinting is done full the combined approach uses more time than fingerprinting alone, but is significantly faster than cross correlation.

4.2. Real experiment

As a final experiment, recordings were done of routine office work, and the combined approach was used to synchronize them. Two PDAs, one placed on the lapel of a student and another placed in the office, recorded about 90 minutes of audio. Some of the elapsed time was spend working in the office (same environment), and the rest of the time consisted of a meeting outside of the office (different environments). The results are shown in figure 8. The decision thresholds from the training session were used. As is evident in the figure, the system is able to classify whether two recordings are from the same environment. The error rate is 0.017 which is very similar to the found test errors. All errors are in the 'same environment' states.

Manual inspection of the found lags was performed as well, but the ground truth is not available, so a performance value is not available. The third plot shows the reported lags and as can be seen they are quite constant indicating that the

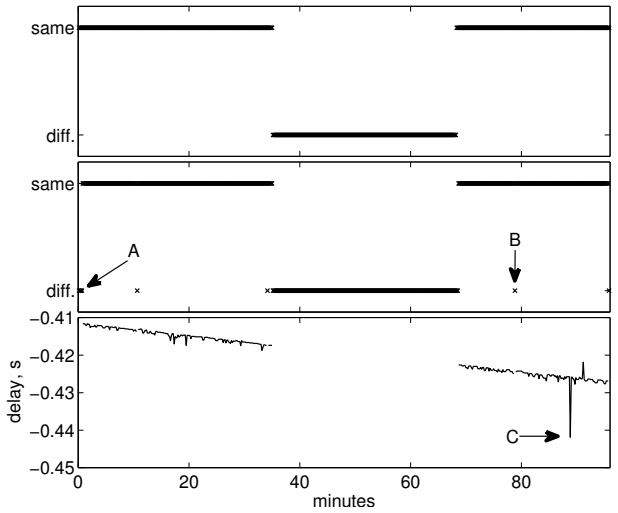

Fig. 8. The combined approach is used to synchronize two recordings. The top figure shows the manually tagged labels, and the second shows the estimated labels. Accurate classification is achieved with an error rate of 0.017. Four errors are found in the beginning of the recording (A) because of handling noise in one recording when attaching the microphone to the lapel. Remaining errors are single points scattered randomly in parts from 'different environments' (B). The third plot shows the estimated delays. A peak is seen in the lags (C) which could be a case of a side peak (cf. figure 2). A decreasing trend is observed indicating that the sample frequencies do not exactly match for the two recording devices, viz. $2.7\frac{\mu s}{s}$ is lost. The used pda's had their clocks synchronized just before starting the experiment and still a delay of $0.4s$ is found. Such delays would have a severe impact on, e.g., a convolutive ICA algorithm further substantiating the need for synchronization.

'true lag' is found. An interesting trend can be observed i.e. the lag changes slightly over time. The reason for this is probably that the clocks in the two devices are not accurate and therefore the sampling rates are slightly different between the two recordings causing a drift.

5. CONCLUSION

In this paper the first steps toward the analysis of multiple Lifelog audio recordings were taken. The steps included clustering recordings into joint environments followed by a synchronization step of recordings within a given audio event group. Two approaches were investigated, showing similar classification performance, but having different advantages. Cross-correlation has sample precision in the found delays and thus can give more accurate synchronization. The fingerprinting method scales much better with the number of recordings. The time complexity is likely to be a serious challenge for real world applications. A joint approach of the two methods was implemented and obtained both the accurate sample resolution and an increased execution speed.

Using the method suggested here the subsequent ICA blind separation problem could be limited to two four-source recordings, instead of one eight-source recordings. For convolutive ICA the shorter the convolutive filters are, the faster and better results are typically obtained. By synchronizing the recordings prior to solving the problem, the filters are limited to only capture the inherent delay from the different distances between microphones and sources. This greatly shortens the length of the filter, from potentially $5 - 10$ s to below 1 s.

An experiment was performed in a real setting and showed that it is indeed possible to detect audio environment similarities and to synchronize the recordings.

6. REFERENCES

[1] Vannevar Bush, "As we may think," *Atlantic Monthly*, July 1945.

[2] Gordon Bell and Jim Gemmell, "A digital life," *Scientific American*, vol. 296, no. 3, pp. 58–65, March 2007.

[3] Eric Freeman and David Gelernter, "Lifestreams: a storage model for personal data," *SIGMOD Rec.*, vol. 25, no. 1, pp. 80–86, March 1996.

[4] Defense Advanced Research Projects Agency (DARPA), "Lifelog: Proposer information pamphlet," 2003.

[5] Datong Chen, Jie Yang, Robert Malkin, and Howard D. Wactlar, "Detecting social interactions of the elderly in a nursing home environment," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, Feb. 2007.

[6] Daniel P. W. Ellis and Keansub Lee, "Accessing minimal-impact personal audio archives," *IEEE Multimedia*, vol. 13, no. 4, pp. 30–8, 2006.

[7] Tanzeem Choudhury and Alex Pentland, "Characterizing social networks using the sociometer," in *North American Association of Computational Social and Organizational Science*, Pittsburg, Pennsylvania, June 2004.

[8] Te-Won Lee, Anthony J. Bell, and Russell H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*, Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, Eds. 1997, vol. 9, p. 758, The MIT Press.

[9] Jean Carletta, "Announcing the AMI meeting corpus," *ELRA Newsletter*, vol. 11, no. 1, pp. 3–5, January 2006.

[10] Avery Wang, "The shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[11] James P. Ogle and Daniel P. W. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 233–236, 2007.

[12] Joe H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.