IDIAP RESEARCH REPORT

IMPACT OF EXCITATION FREQUENCY ON SHORT-TERM RECORDING SYNCHRONISATION AND CONFIDENCE ESTIMATION

Danil Korchagin

Idiap-RR-20-2011

JUNE 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

IMPACT OF EXCITATION FREQUENCY ON SHORT-TERM RECORDING SYNCHRONISATION AND CONFIDENCE ESTIMATION

Danil Korchagin

Idiap Research Institute
Martigny, Switzerland

ABSTRACT

In this paper, we present the results of a study on excitation frequency impact on short-term recording synchronisation and confidence estimation for multisource audiovisual data, recorded by different personal capturing devices during social events. The core of the algorithm is based on perceptual time-quefrency analysis with a precision of 10 ms. Performance levels achieved to date on 14+ hours of hand-labelled dataset have shown positive impact of excitation frequency on temporal synchronisation (98.19% precision for 5 s recordings) and confidence estimation (99.08% precision with 100% recall for 5 s recordings). The results surpass the performance of fast cross correlation while keeping lower system requirements.

1. INTRODUCTION

The TA2 project (Together Anywhere, Together Anytime) [1] is concerned with investigation of how multimedia devices can be introduced into a family scenario to break down technology and distance barriers. Technically, the TA2 project tries to improve group-to-group communication by making it more natural and by giving the users the means to easily participate in shared activities. In this sense, we are interested in the use of consumer level multimedia devices in novel application scenarios.

One generic scenario is the use of multiple capture devices at the same event with subsequent navigation through captured data within a common timeline (see figure 1). In a professional scenario, one might expect to be able to use multiple capture devices, and for them all to be synchronised via a common clock or similar [2]. Consumer level devices, however, do not normally provide such capabilities and are turned on and off at the will of their users. This leaves us with the metadata and audiovisual signals to infer synchronisation information. The available camera time and recording time in the metadata are based on the personal capturing devices and are most likely to be different across the devices. Another intuitive and simple approach would be to compare the corresponding audiovisual signals. However, recordings captured at the same time by different cameras may look and sound different because of camera locations (e.g. different lighting ambience, noisy surrounding), camera settings (e.g. white point balance, audio gain), quality of the camera components (e.g. sensor, lens, microphones). Therefore, raw audiovisual signals are not suitable for synchronisation purpose.

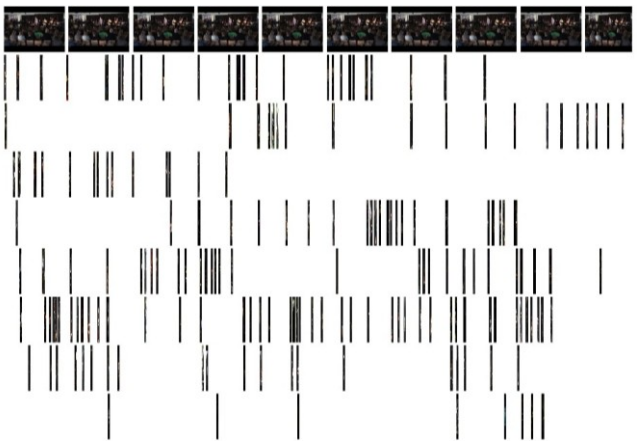

Figure 1 – An example of automatic synchronisation of 187 test signals 5 s each within an event of 1 h 37 min.

The solution would be to automatically synchronise the multisource recordings by detecting and matching audio and/or video features extracted from the content.

Early studies on video-based synchronisation techniques [3, 4] relied on assumptions of static cameras and homographic images. In [5] a usage of tracking a line feature in multiple videos with limited camera motion is used, though the method implies identical frame rate on all cameras. In [6] moving features are computed that best relate with the pre-computed camera geometries, nevertheless the method depends on sufficient texture for tracking and other constraints. In [7] authors propose a synchronisation based on flash sequences, which is suitable only for particular type of events. Other state of the art video-based synchronisation techniques [8, 9, 10, 11] also impose controlled environments. Therefore, if the devices are hand-held and environments are unconstrained, we cannot rely in any predictable sense on the video signal. This leaves us with the audio signal from which to infer synchronisation information.

One of audio-based solutions is the use of audio onsets [12], which are the perceived starting points in an auditory event. Many other solutions rely on audio fingerprinting techniques [13, 14, 15, 16], which result in fairly good but not perfect synchronisation of the recordings.

In our previous study [17] we have shown that the auxiliary signals can be synchronised with the reference signal reliably based on audio features typical of ASR applications. The present investigation concerns further study in the direction of excitation frequency impact on short-term recording synchronisation and confidence estimation.

2. SHORT-TERM SYNCHRONISATION

Consider a music performance. The duration of the corresponding event can easily be of the order of a small number of hours. It is normal in such situations to decrease the search space, retaining only useful information for synchronisation. In our previous study [17] we have shown that multiple recordings can be synchronised to an acceptable accuracy using audio features typical of ASR applications and corresponding confidence can be reliably estimated. For recordings longer than 15 s we were able to achieve 100% precision on 100 recording dataset for time-quefrency signatures without excitation frequency versus 98% for fast cross correlation. For recordings shorter than 5 s the precision levels were lower due limited length of the signatures and the real world variability of the data (noise, reverberation, non-stationarity of cameras, etc).

In our study we define the recordings not longer than 5 s as short-term recordings. In following chapter we present the experimental results of a study on excitation frequency impact achieved to date on 14+ hours of hand-labelled dataset. This is achieved via redefinition of time-quefrency signatures as described below. The re-estimated precision dependency on the length of test signals in respect to enlarged dataset (997 test recordings) is shown in figure 2.

We define time-quefrency signatures as time-quefrency matrices based on normalised truncated mel-cepstral vectors in steps of 10 ms. A 256 point Discrete Fourier Transform (DFT) is performed on overlapping audio frames of 16 ms in steps of 10 ms and squared to give the power spectrum. The resulting 129 unique bins are then decimated using a filter-bank of 23 overlapping triangular filters equally spaced on the mel-scale. The mel-scale corresponds roughly to the response of the human ear. A logarithm and DFT then yield the mel-cepstrum [18]. Lower 13 dimensions retain the energy and general spectral shape, while higher dimensions retain excitation frequency [19], which is normally truncated. In this study we keep higher mel-cepstrum coefficients, related to excitation frequency, to estimate corresponding impact on short-term recording synchronisation and confidence estimation. The energy is truncated for proposed approach, though kept for a subset of other considered signatures. Next, Cepstral Mean Normalisation (CMN) is performed by subtracting from each cepstral vector the mean of the vectors of the preceding (approximately) half second. This has the effect of removing convolutional channel effects. Finally, if the norm of a vector of the mean normalised cepstral coefficients is higher than 1, then the vector is normalised in Euclidean space. This gives us the reduced variance of the search distance space.

Synchronisation, based on the above time-quefrency signatures, is performed by searching for a best distance [17] in n-dimensional Euclidean space between the time-quefrency representations H_j^i and G_j of test and reference signals h_j^i and g_j, the relative position within the signal g_j is given by:

$$t_j^i = \alpha \cdot \arg\min_{G_j}(d(H_j^i, G_j)),$$

where d is Euclidean metric, α is the step within time-quefrency representation in s.

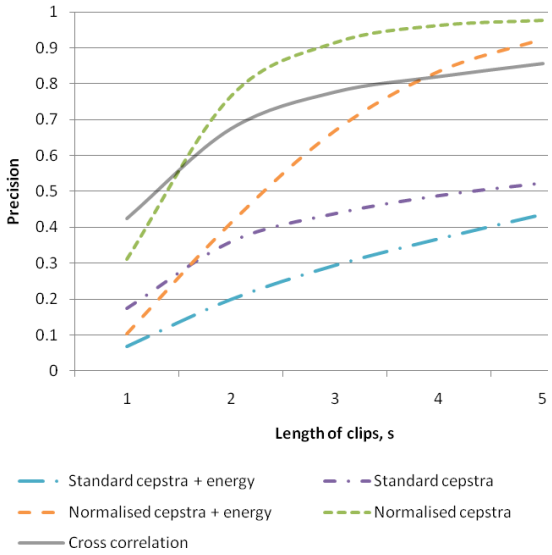

Figure 2 – Precision versus test signal length.

In the case of fast cross correlation, the relative position within the signal g_j is given by:

$$t_j^i = \frac{1}{f_s}\arg\max\left(F^{-1}\left(\left(F\{h_j^i\}\right)^* \cdot F\{g_j\}\right)\right)$$

In the above formulation, the parameters are as before. F denotes the fast Fourier transform. f_s is the sampling frequency. An asterisk indicates the complex conjugate. Cross correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. It is well known technique and can be used to search a long duration signal for a shorter.

The confidence of the above techniques can be estimated as a measure of relative variance of the search space via standard deviation. For time-quefrency signature based technique, the standard deviation can be replaced by the maximum distance [17]. Thus the confidence can be estimated by searching for a confidence corresponding to a best distance in n-dimensional Euclidean space between time-quefrency representation of test and reference signals:

$$C_j^i = \frac{\left|\mathrm{E}(d_{G_j}^{(i,j)}) - \min_{G_j}(d_{G_j}^{(i,j)})\right|}{4 \cdot \left|\mathrm{E}(d_{G_j}^{(i,j)}) - \max_{G_j}(d_{G_j}^{(i,j)})\right|} - 0.2 l_{(i,j)}^{-1}$$

In the above equation, C_j^i is the confidence measure of successful synchronisation of test and reference signals h_j^i and g_j. E denotes the expectation. $l_{(i,j)}$ is the length of test signal h_j^i in s.

In the case of fast cross correlation, the confidence estimation is given by [20]:

$$C_i^j = \frac{1}{40\sigma}\max\left(F^{-1}\left(\left(F\{h_i^j\}\right)^* \cdot F\{g_j\}\right)\right),$$

where σ is the standard deviation of the cross correlation.

It is worth mentioning that the use of standard cross correlation instead of fast cross correlation is not feasible as it is computationally onerous (several days per test signal instead of few minutes on an Intel Core 2 CPU 6700 2.66GHz).

3. EXPERIMENTAL RESULTS

All results presented in this paper were achieved on a real life dataset of 1010 recordings:

- 13 reference signals (total length – 13 h 31 min), recorded with:
  - Canon XL-G1,
  - Sony HDR-520VE,
  - Benq-Siemens E71.
- 997 test signals of 5 s each (total length – 1 h 23 min), recorded with:
  - Nokia N95,
  - Canon FS100E mini,
  - Canon XM1 mini DV,
  - Sony DCR-PC3e,
  - Sanyo Xacti HD mini,
  - iPhone 3G S,
  - Canon Powershot S5IS,
  - Panasonic Lumix DMC-LX3,
  - Sony PDC-100E,
  - Panasonic Lumix DMC-FX500,
  - Sony PDC-10E,
  - Nikon D70,
  - Panasonic Lumix DMC-F57,
  - Fujifilm camera,
  - Sony DSC-V1,
  - Sony Ericsson G502,
  - etc.

The recordings were captured by several social groups of people (with up to 12 socially connected people per group) during 13 different events in 3 different countries within Europe. The reference signal contents consist of musical concerts/rehearsals with multiple events/replays one after the other. All corresponding audio tracks were extracted and converted to 16 kHz mono PCM files with FFMPEG software [21].

Experiments were conducted on a closed set (i.e. we did not consider test signals that did not correspond to the reference signal). Nevertheless according to our previous study on a rejection mechanism [20], the proposed approach can be successfully extended to an open set.

To avoid possible inaccuracy associated with manual annotation (the ear is insensitive to delays below 160 ms) and limited speed of sound (each 10 m distance from the object results in 1 frame lag) the precision was calculated as the number of correctly (within ±5 frames) synchronised clips divided by the total number of test clips. This is a bit wider range than ITU-R recommendation [22], proposing the range between -125 ms and +45 ms as a requirement for editing multiple recordings without losing lip synchronisation. While theoretically it is feasible to reduce our experimental range for scoring to fit ITU-R recommendation, it would require a lot of additional work to update annotations in respect to required ITU-R accuracy.

In figure 3 we illustrate how the dimensionality of the feature vector including excitation frequency range influences precision of short-term recording synchronisation.

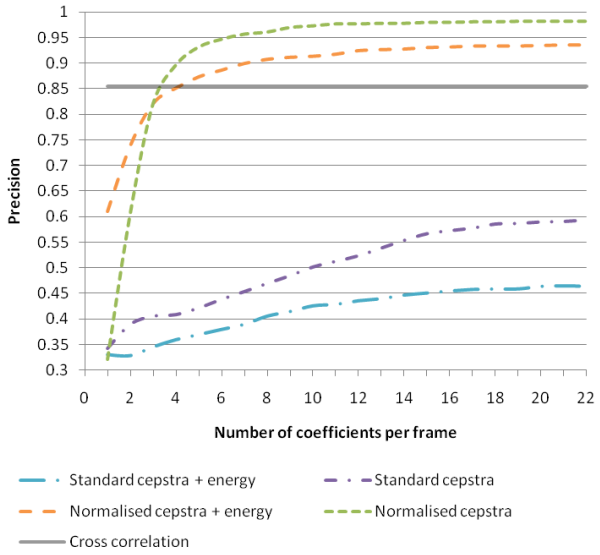

Figure 3 – Precision versus number of coefficients.

It is clearly visible that the precision improves with increasing cepstral analysis order. Precision for lower 12 dimensions, corresponding to the general spectral shape, results in 97.69%, while additional 7-10 coefficients, corresponding to excitation frequency range, allows to increase precision level of synchronisation up to 98.19%. I.e. we observe absolute improvement by 0.5% in the case of excitation frequency use. This in turn corresponds to 21.6% relative improvement in respect to error rate achieved on described dataset and based on the technique from our previous study [17] (from 2.31% to 1.81%). However, precision is lower when the energy is considered or normalisation in Euclidean space is excluded. We hypothesise this is due to the increased variance of the search distance space.

In figure 4 we illustrate how the dimensionality of the feature vector including excitation frequency range influences confidence estimation distribution. Here we consider only the case when energy is excluded and normalisation in Euclidean space is applied. The graph contains in total 21’934 confidence estimates for both positive (green dots) and negative (red dots) classes of synchronisation. The positive class is defined as set of test signals, properly synchronised with the reference signal. The negative class is defined as set of test signals, misaligned with the reference signal. While we observe positive impact of excitation frequency on reducing negative class, we have to state that corresponding negative class is becoming wider and sparser. Also an optimal separation of positive and negative classes for short-term recordings is much trickier than if we would have the recordings of 30+ s. One of generic solution for this two class classification problem would be the use of machine learning approach, e.g. the support vector machine [23]. Nevertheless, depending on subsequent application, the weights for corresponding classes can be different. This is why, it is important to know not only confidence estimates distribution, but the dependency between precision and recall values.

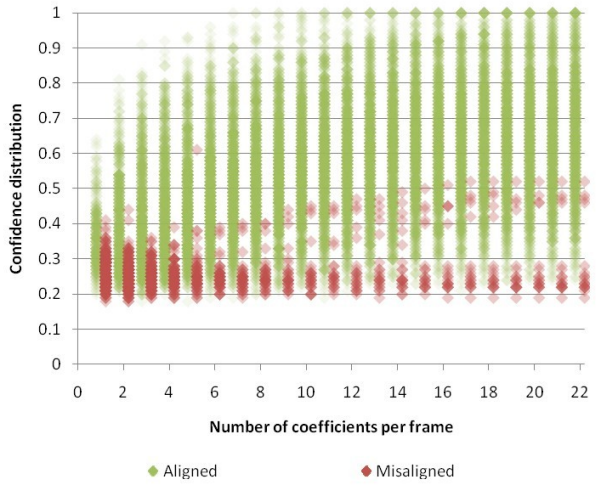

Figure 4 – Confidence distribution versus number of coefficients.

Dependency between precision and recall values can be estimated experimentally via application different confidence threshold values. In Figure 5 this dependency is illustrated for 9 selected cases. Precision is defined as the number of true positive test signals (test signals correctly detected as belonging to the positive class) divided by the total number of test signals detected as belonging to the positive class (the sum of true positive and false positive test segments). Recall is defined as the number of true positives test signals divided by the total number of test signals that actually belongs to the positive class (the sum of true positive and false negative test signals). Prefix "standard" means no normalisation in Euclidean space is performed. Prefix "normalized" denotes normalisation in Euclidean space is performed. Signatures with lower 12 dimensions, corresponding to the general spectral shape, are marked as "cepstra". Signatures with lower 22 dimensions, corresponding to the general spectral shape and excitation frequency, are marked as "cepstra + excitation". Signatures with lower 12 dimensions and energy are marked as "cepstra + energy". Signatures with lower 22 dimensions and energy are marked as "cepstra + energy + excitation". To allow better positioning with other techniques we present the results for well-known fast cross correlation method as well.

It is clearly visible, that 4 out of 8 time-quefrency signature based techniques for confidence estimation perform better than confidence estimation based on fast cross correlation. The best result belongs to the case when the general spectral shape is combined with excitation frequency and normalised in Euclidian space (double square dot green line). We were able to achieve 99.08% precision (versus 97.79% for the general spectral shape without excitation frequency) in the case of 100% recall and 76.00% recall (versus 75.46% for the general spectral shape without excitation frequency) in the case of 100% precision for confidence estimation. It is worth mentioning not perfect smoothness of the graphs. We suppose this is due to the limited amount of the test signals and better smoothness might be obtained by enlargement of test dataset by factor of 10.

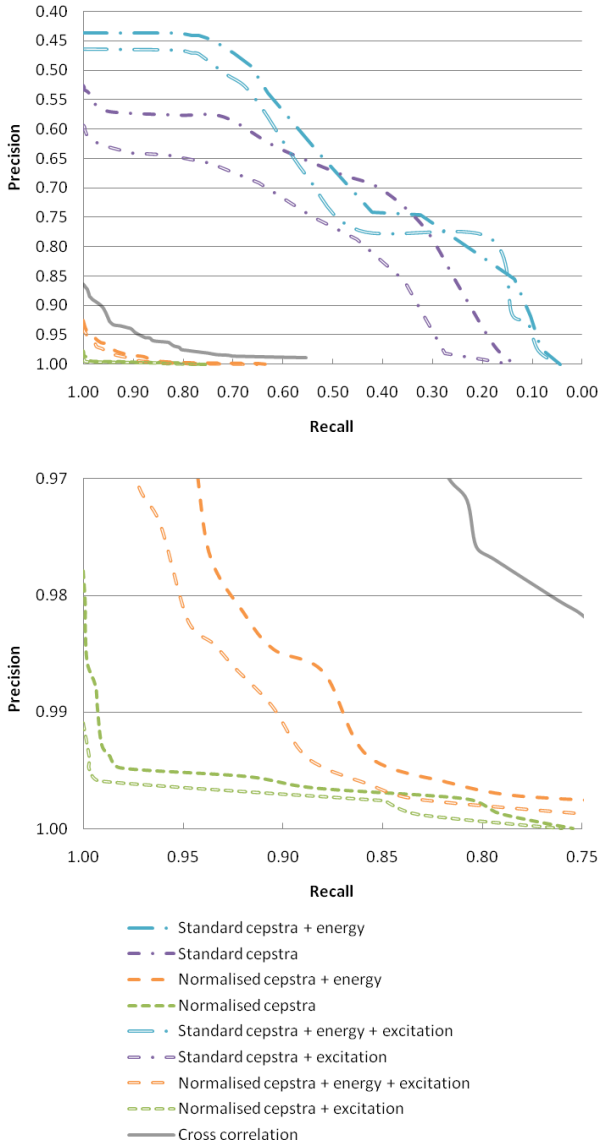

Figure 5 – Precision versus recall for confidence estimation.

Processing time (on an Intel Core 2 CPU 6700 2.66GHz) for the proposed algorithm without multi-core optimisation was 25 seconds for automatic synchronisation of a 5 second test signal over the 51 min reference signal using the general spectral shape and excitation frequency, 14 seconds for the same test signal using the general spectral shape only, and 70 seconds for the same test signal using fast cross correlation technique. It is directly proportional to the length of the test signal, to the length of the reference signal and to the feature vector dimensionality. Thus we can conclude that computational efficiency of proposed approach is even better than fast cross correlation and memory requirement is about 28% of the size of reference signal (28 MB versus 3 GB for fast cross-correlation). There is clearly a trade-off between desirable precision/recall levels and execution time / memory requirements. By lowering the cepstral order we can surely reduce execution time, memory requirements, and precision/recall levels.

4. CONCLUSION

We have shown the positive impact of excitation frequency on short-term recording synchronisation and confidence estimation. We have confirmed generalization of the results on 14+ hours of hand-labelled dataset. We have estimated that the energy of the signal is not good for synchronisation even when excitation frequency is considered. We have estimated dependencies between precision and recall levels for confidence estimation. We have shown that results surpass the precision and recall levels of fast cross correlation, while keeping lower system requirements.

5. ACKHOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme ICT Integrating Project “Together Anywhere, Together Anytime” (TA2, FP7/2007-2013) under grant agreement no. ICT-2007-214793. We are grateful to British Telecom and Centrum Wiskunde & Informatica for provision of additional sessions for the real life dataset.

REFERENCES

[1] Integrating project within the European research programme 7, "Together anywhere, together anytime", http://www.ta2-project.eu, 2008.

[2] J.-M. Verrier, "Audio boards and video synchronisation", in *Proceedings of the AES UK 14th Conference: Audio - The Second Century*, London, UK, 1999.

[3] G. P. Stein, "Tracking from multiple view points: self calibration of space and time", in *Proceedings of the DARPA IU Workshop*, pp. 521–527, 1998.

[4] Y. Caspi, D. Simakov, and M. Irani, "Feature based sequence-to-sequence matching", in *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, 2004.

[5] C. Lei and Y. H. Yang, "Tri-focal tensor based multiple video synchronization with sub-frame optimization", in *IEEE Trans. on Image Processing*, 2005.

[6] A. Whitehead, R. Laganire, and P. Bose, "Temporal synchronization of video sequences in theory and in practice", in *Proceedings of the 14th Brazilian Symposium on Computer Graphics and Image Processing*, pp. 132–137, 2005.

[7] P. Shrestha, H. Weda, M. Barbieri, and D. Sekulovski, "Synchronization of multiple videos using still camera flashes", in *Proceedings of the 14th ACM International Conference on Multimedia*, pp. 137–140, 2006.

[8] Y. Caspi and M. Irani, "Aligning non-overlapping sequences", in *International Journal of Computer Vision*, vol. 48, n. 1, 39–51, 2002.

[9] W. Yan and M. S. Kankanhalli, "Detection and removal of lighting & shaking artefacts in home videos", in *Proceedings of the 10th ACM international conference on Multimedia*, pp. 107–116, 2002.

[10] S. N. Sinha and M. Pollefeys, "Visual-hull reconstruction from uncalibrated and unsynchronized video streams", in *Proceedings of the 3D Data Processing, Visualization, and Transmission*, 2nd International Symposium, 2004.

[11] T. Tuytelaars and L. Van Gool, "Synchronizing video sequences", in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[12] J. P. Bello, L. Daudet, S. Abdallah, et al., "A tutorial on onset detection in music signals", in *IEEE Transactions on Speech and Audio Processing*, vol. 13, issue 5, part 2, 2005.

[13] M. Cremer and R. Cook, "Machine-assisted editing of user generated content", in *Proceedings of SPIE-IS&T Electronic Imaging*, vol. 7254, 2009.

[14] L. Kennedy and M. Naaman, "Less talk, more rock: automated organization of community-contributed collections of concert videos", in *Proceedings of the 18th ACM International Conference on World Wide Web*, pp. 311–320, 2009.

[15] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system", in *Proceedings of the International Symposium on Music Information Retrieval*, 2002.

[16] P. Shrestha, H. Weda, and M. Barbieri, "Synchronization of multi-camera video recordings based on audio", in *Proceedings of the 15th annual ACM International Conference on Multimedia*, 545–548, 2007.

[17] D. Korchagin, P. N. Garner, and J. Dines, "Automatic temporal alignment of AV data with confidence estimation", in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, 2010.

[18] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental", in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., pp. 374-388, Academic, New York, USA, 1976.

[19] L. Rabiner, B.-H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, Upper Saddle River, NJ, USA, 1993.

[20] D. Korchagin, "Out-of-scene AV data detection", in *Proceedings IADIS International Conference on Applied Computing*, vol. 2, pp. 244-248, Rome, Italy, 2009.

[21] Open source multiformat multimedia conversion tool "FFMPEG", http://www.ffmpeg.org.

[22] BT.1359-1, "Relative timing of sound and vision for broadcasting", in *Recommendation ITU-R*, 1998.

[23] V. N. Vapnik, *The nature of statistical learning theory*, Springer, 2nd edition, 2000.