AUTOMATIC TEMPORAL ALIGNMENT OF AV DATA WITH CONFIDENCE ESTIMATION

Danil Korchagin, Philip N. Garner and John Dines

Idiap Research Institute
CH-1920 Martigny, Switzerland

ABSTRACT

In this paper, we propose a new approach for the automatic audio-based temporal alignment with confidence estimation of audio-visual data, recorded by different cameras, camcorders or mobile phones during social events. All recorded data is temporally aligned based on ASR-related features with a common master track, recorded by a reference camera, and the corresponding confidence of alignment is estimated. The core of the algorithm is based on perceptual time-frequency analysis with a precision of 10 ms. The results show correct alignment in 99% of cases for a real life dataset and surpass the performance of cross correlation while keeping lower system requirements.

Index Terms — time-frequency analysis, time synchronisation, pattern matching, reliability estimation

1. INTRODUCTION

The TA2 project (Together Anywhere, Together Anytime) is concerned with investigation of how multimedia devices can be introduced into a family scenario to break down technology and distance barriers. 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 in a single room. The present investigation concerns the possibility of using multiple consumer level video cameras to reconstruct the narrative of the recorded event and to align different sources. 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 [1]. Consumer level devices, however, do not normally provide such capabilities. Further, if the devices are hand-held, we cannot rely in any predictable sense on the video signal. This leaves us with the audio signal [2], [3] from which to infer synchronisation information.

In this study, we were provided with a single reference signal from a fixed camera that recorded the whole scene. We were also provided with several auxiliary signals from hand-held cameras that recorded parts of the scene. If we could show that the auxiliary signals could be aligned with the reference signal reliably, then the project could profit from using audio-based temporal alignment. If it were too error-prone or computationally onerous, then other solutions would have to be sought.

2. EXPERIMENTAL DATASET

All results presented in this paper were achieved on a real life dataset of 100 recordings:

Source	Length range	Audio spec.	Video spec.
Canon camera XL-G1 SD/HD (master track)	51 min	PCM 32000Hz Stereo 1024Kbps	DVSD 720x576 25.00fps 28799Kbps
Smartphone Nokia N95 (17 clips)	12-130 s	AAC 48000Hz Mono 768Kbps	MPEG4 640x480 28.60fps 2700Kbps
Canon camera FS100 mini (28 clips)	4-133 s	Dolby AC3 48000Hz Stereo 256Kbps	MPEG2 720x576 25.00fps 9600Kbps
Sony camera DCR-PC3e (15 clips)	16-695 s	PCM 48000Hz stereo 1536Kbps	DVSD 720x576 25.00fps 28800Kbps
Sanyo camera Xacti HD mini (39 clips)	1-250 s	AAC 48000Hz Stereo 1536Kbps	H.264 1280x720 29.97fps 5975Kbps

The master track content consists of a high school rehearsal with multiple events/replays one after the other. All corresponding audio tracks were extracted and converted to 16 kHz mono PCM files with VirtualDub software [4].

Experiments were conducted on a closed set (i.e. we did not consider a rejection mechanism for test segments that did not correspond to the master track). Nevertheless according to our previous studies on a rejection mechanism [5], the proposed approach can be successfully extended to an open set.

3. TEMPORAL ALIGNMENT

Consider a simple high school concert event. The duration of the corresponding master track can easily be of the order of a small number of hours. This in turn corresponds to a large quantity of raw audio data (stereo at 48 kHz). It is normal in such situations to decrease the search space, retaining only useful information for temporal alignment. Accordingly, we assume from the outset that raw PCM audio data is both too voluminous and too noisy to produce good audio alignment. We suppose that good results might be obtained by lower resolution features such as frame energy and cepstra. Certainly, a resolution approaching video frame rate is sufficient for the purposes of our application.

Given that our broader application is expected to include Automatic Speech Recognition (ASR), the pre-processing takes the form of a standard feature extraction chain used in ASR. In our work we use Mel Frequency Cepstral Coefficients (MFCC) [6] with a 10 ms frame rate. MFCC is a perceptually motivated spectrum representation that is widely used not only in speech recognition but also for music modelling [7]. Such pre-processing includes energy-like features (actually the zero'th cepstral coefficient) along with cepstra representing the general spectral shape.

We assume that test samples are relatively short, thus we can ignore the clock skew problem between test and reference (i.e., there is almost zero skew due to unsynchronised clocking of different devices). Presumably in some cases for long recordings the two could become misaligned, in which case additional techniques such as dynamic time warping [8] should be taken into account during the matching process. We consider two operating modes, one the well-known cross correlation and the other template matching based on ASR-related features.

3.1. Cross correlation

Cross correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. It can be used to search a long duration signal for a shorter. If h_i and g are the raw test and reference signals respectively, and h^* is the complex conjugate of h, then t_i, the relative position in ms of the i'th test clip, is given by:

$$t_i = \frac{10^3}{f_s} \arg\max_n \left(\sum_m h_i^*(m) g(n+m) \right),$$

where f_s is the sampling frequency.

Regardless of the simplicity of implementation, standard cross correlation cannot be implied by our scenario as it is computationally onerous (several days per clip on an Intel Core 2 CPU 6700 2.66GHz), nevertheless this can be resolved by the convolution theorem and fast Fourier transform, also known as fast cross correlation:

$$t_i = \frac{10^3}{f_s} \arg\max \left(F^{-1} \left(\left(F\{h_i\} \right)^* \cdot F\{g\} \right) \right)$$

In the above formulation, the parameters are as before, except F denotes the fast Fourier transform. An asterisk again indicates the complex conjugate. The processing time for fast cross correlation takes only 70 seconds per clip, though it requires much more RAM (3 GB versus 100 MB).

3.2. Template matching

Audio is down-sampled (if necessary) to 16 kHz and pre-emphasised to flatten the spectral shape. A 256 point Discrete Fourier Transform (DFT) is performed 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, which is truncated, retaining the lower 13 dimensions. This truncation retains spectral shape and discards excitation frequency. 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, the 13 normalised cepstral coefficients are then augmented by first and second order derivatives, corresponding to their velocity and acceleration. This gives $k=39$ dimensional vectors.

Template matching based on the above features is performed by searching for a best distance in n-dimensional Euclidean space between the test time-quefrency matrix (corresponding to a test clip) and the master time-quefrency matrix in steps of 10 ms. If V_i is the i'th test matrix ($1 \leq i \leq S$, where S is the number of test clips), M is the master matrix and $M_p^{(i)}$ is the sub-matrix of the master matrix, shifted from the beginning by $10p$ ms, then t_i, the relative position in ms of the i'th test matrix, is given by:

$$t_i = 10 \cdot \arg\min_{M_p^{(i)} \in M} \left(d(M_p^{(i)}, V_i) \right),$$

where the metric d is given by:

$$d(M_p^{(i)}, V_i) = \sum_{q=1}^{N_i} \| \alpha_{p+q} m_{p+q} - \beta_{i,q} v_{i,q} \|$$

In the above equation, $v_{i,q}$ is the k-dimensional vector of the i'th test matrix, which corresponds to a frame q represented by k pre-processed coefficients. N_i is the number of frames inside the matrix V_i. m_p is the k-dimensional vector of the master matrix shifted from the beginning by $10p$ ms and corresponding to a frame p represented by k pre-processed coefficients. α_{p+q} and $\beta_{i,q}$ are normalisation coefficients for corresponding frames $p+q$ and q of the matrices M and V_i:

$$\alpha_{p+q} = \begin{cases} 1\,, if \parallel m_{p+q} \parallel \leq 1 \\ \dfrac{1}{\parallel m_{p+q} \parallel}\,, if \parallel m_{p+q} \parallel > 1 \end{cases}$$

$$\beta_{i,q} = \begin{cases} 1\,, if \parallel v_{i,q} \parallel \leq 1 \\ \dfrac{1}{\parallel v_{i,q} \parallel}\,, if \parallel v_{i,q} \parallel > 1 \end{cases}$$

The elements α_{p+q} and $\beta_{i,q}$ are upper thresholded at 1 to decrease the impact of quiet frames. (In standard mode they are fixed at 1.)

The dimension n of the search space is equal to the length of the master track in steps of 10 ms.

3.3. Experimental results

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 performance was calculated as the number of correctly (within ±5 frames) aligned clips divided by the total number of test clips.

In figure 1 we illustrate how the dimensionality of the feature vector influences total performance.

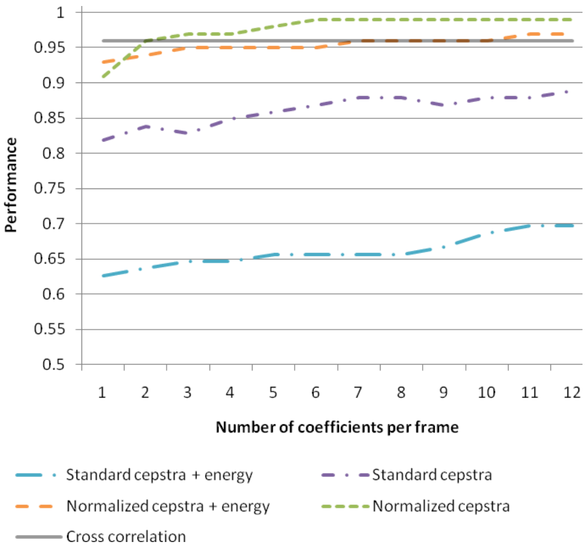

Figure 1. Performance versus number of coefficients.

It is clearly visible that the performance improves with increasing cepstral analysis order. However, it is dramatically lower when the energy is considered (dash dot and long dash dot lines). We hypothese this is due to the increased variance of the search distance space. Delta features do not have any influence (and thus were excluded from the figure), we hypothesise due to the fact that deltas can be easily reconstructed from cepstra over time. Further, they are used in ASR as a continuity constraint, which is not necessary in this application. Normalisation (via the elements α_{p+q} and $\beta_{i,q}$) allows to surpass the performance of cross correlation and results in 99% versus 96% for cross correlation, we believe due to the reduced variance of the normalised search distance space.

Nevertheless, there is also a strong dependency on the length of test recordings. In figure 2 we illustrate how the length of the test segments impacts on the total performance. The performance grows and, for recordings longer than 15 s, 100% performance is achievable on the described dataset for the proposed approach versus 98% for cross correlation. For recordings shorter than 5 s the difference between the performance of the proposed approach and the performance of cross correlation varies up to 12%. We suppose that alignment of very short recordings is not robust due the real world variability of the data (noise, reverberation, non-stationarity of cameras, inter-microphone variability, inter-codec variability, etc).

It is worth mentioning that the variance of the search space is directly proportional to the length of the test recordings. This is why, on long recordings, we observe quite good results even for standard cepstra.

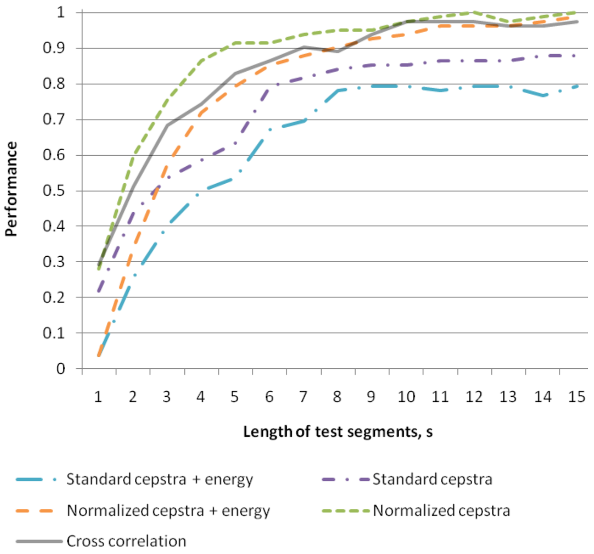

Figure 2. Performance versus test segment length.

Processing time (on an Intel Core 2 CPU 6700 2.66GHz) for the proposed algorithm without multi-core optimisation was 14 seconds for automatic temporal alignment of a 12 second test recording over the 51 min master track using 5 cepstra, and 33 seconds for the same segments using 12 cepstra. It is directly proportional to the length of the test segment, to the length of the master track 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 15% of the size of reference signal (15 MB versus 3 GB for fast cross-correlation).

4. CONFIDENCE ESTIMATION

The confidence of the above techniques can be estimated as a measure of relative variance of the search space via standard deviation. For template matching based on ASR-related features, the standard deviation can be replaced by the maximum distance. Thus the confidence estimation is performed by searching for a confidence corresponding to a best distance in *n*-dimensional Euclidean space between test time-quefrency matrix (corresponding to a test clip) and master time-quefrency matrix with 10 ms step:

$$C_i = \frac{\left|\mathrm{E}(d_p^{(i)}) - \min_p(d_p^{(i)})\right|}{4 \cdot \left|\mathrm{E}(d_p^{(i)}) - \max_p(d_p^{(i)})\right|} - \frac{20}{N_i}$$

In the above equation, C_i is the confidence measure of matching the *i*'th test matrix and E is expectation. N_i is the number of frames inside test matrix.

In figure 3 we illustrate how the length of the test segment influences the confidence measure. It is worth mentioning that the use of maximum distance instead of standard deviation provides almost the same result, nevertheless requires only 1 pass instead of 2, and thus gives us a speed optimisation of about 2 times.

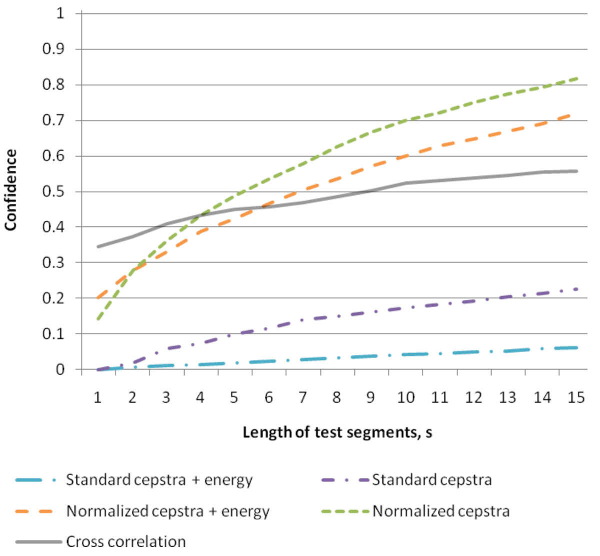

Figure 3. Confidence versus test segment length.

It is clearly visible that the confidence increases with increasing the length of test segments. Additional investigations [5] have proved good robustness of proposed confidence measure and results in 100% of confidence performance for any data with confidence higher than 50%.

5. CONCLUSION

We have shown that multiple AV signals can be aligned to an acceptable accuracy using audio features typical of ASR applications and corresponding confidence can be reliably estimated. Surprisingly, we found that the energy of the signal is not good for alignment, but that good alignment can be inferred from a small number of normalised cepstra. We have shown that results can be improved using a feature vector normalisation and surpass the performance of fast cross correlation, while requiring less resources.

6. ACKNOWLEDGMENTS

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 for provision of the real life dataset.

7. REFERENCES

[1] Verrier, J.-M., "Audio Boards and Video Synchronisation", *Proceedings of the AES UK 14th Conference: Audio - The Second Century*, London, UK, 1999.

[2] Dannenberg, R.B. and Hu, N., "Polyphonic Audio Matching for Score Following and Intelligent Audio Editors", *Proceedings of the 2003 International Computer Music Conference*, pp. 27-34, San Francisco, USA, 2003.

[3] Birmingham, W.P., et al., "MUSART: Music Retrieval via Aural Queries", *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR)*, pp. 73-81, Bloomington, USA, 2001.

[4] Open source video capture and processing program VirtualDub, http://www.virtualdub.org/

[5] Korchagin, D., "Out-of-Scene AV Data Detection", *Proceedings IADIS International Conference on Applied Computing*, vol. 2, pp. 244-248, Rome, Italy, 2009.

[6] Mermelstein, P., "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.

[7] Logan, B., "Mel Frequency Cepstral Coefficients for Music Modeling", *Proceedings of the 1st International Symposium on Music Information Retrieval*, Plymouth, USA, 2000.

[8] Hu, N., Dannenberg, R.B. and Tzanetakis, G., "Polyphonic Audio Matching and Alignment for Music Retrieval", *In 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 185-188, New York, USA, 2003.