Memoirs of Togetherness from Audio Logs

Danil Korchagin

Idiap Research Institute,
P.O. Box 592, CH-1920 Martigny, Switzerland
Danil.Korchagin@idiap.ch

Abstract. In this paper, we propose a new concept how tempo-social information about moments of togetherness within a social group of people can be retrieved in the palm of the hand from social context. The social context is digitised by audio logging of the same user centric device such as mobile phone. Being asynchronously driven it allows automatically logging social events with involved parties and thus helps to feel at home anywhere anytime and to nurture user to group relationships. The core of the algorithm is based on perceptual time-frequency analysis via confidence estimate of dynamic cepstral pattern matching between audio logs of people within a social group. The results show robust retrieval and surpass the performance of cross correlation while keeping lower system requirements.

Keywords: Time-frequency analysis, pattern matching, confidence estimation.

1 Introduction

The TA2 project (Together Anywhere, Together Anytime) is concerned with investigation of how multimedia devices can be introduced into a family environment to break down technology, distance and time barriers. How can we feel at home in a world where millions of people are in continual movement all around the world? How can we help to nurture social relationships? This is something that current technology does not address well: modern media and communications serve individuals best, with phones, computers and electronic devices tending to be user centric and providing individual experiences. In this sense, we are interested in breaking down the barrier between user centric and group centric media, in creation of mobile domesticity which can automatically generate memoirs of social interactions and fill the gap between user centric media devices and social networks.

Many of our enduring experiences, holidays, festivals, celebrations, concerts and moments of fun are kept as social memoirs. Additional media about these memoirs can be easily retrieved via services-on-demand from social networks. How can we automatically filter out and map only relevant information for personal memoirs of togetherness?

Nowadays more and more users start to use audio logging available in many palm devices. Can we profit from audio logs to augment a distributed domestic space with memoirs of social interactions? Most of the people do not intend to disclose private information and the purpose of each of audio log is primarily personal.

The present investigation concerns the possibility of multiple audio log (recorded by user centric devices such as mobile phones and camcoders) synchronisation for automatic generation of memoirs of togetherness for personal archives. User centric devices do not normally provide such functionality. Further, if people do not share the same acoustic field or the devices are used inside big buildings, the GPS information cannot be used to reliably log social interactions. This leaves us with the audio signal [1] from which to infer a social context [2].

2 Audio Log Processing

Audio logs from user centric devices can be up to 24 hours per day. It is normal in such situations to reduce the initial large quantity of raw audio data, retaining only useful information about social context. In our work we use Mel Frequency Cepstral Coefficients (MFCC) [3] with a 10 ms frame rate. MFCC is a perceptually motivated spectrum representation that is widely used in acoustic and speech processing.

Audio logs are resampled to mono 16 kHz and then processed in pairs within the group of socially connected people on per day basis. External audio log in each pair is divided into subsamples of 30 seconds length each. This has the effect of removing a clock skew problem between different devices (within possible 0.03% range). Presumably the long subsamples could become misaligned, in which case additional techniques such as dynamic time warping [4] should be taken into account. Though some information can be retrieved via high-level modelling [5], we consider only low-level operating modes, one the well-known cross correlation and the other pattern matching based on ASR-related features [6].

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. Corresponding confidence is taken as a proportion between maximum of cross correlation product and its standard deviation. To get real-time computational efficiency we apply the convolution theorem and the fast Fourier transform, also known as fast cross correlation.

In case of pattern matching based on ASR-related features audio is pre-emphasised to flatten the spectral envelope and 13 Mel Frequency Cepstral Coefficients are retrieved in steps of 10 ms. The mel-scale corresponds roughly to the response of the human ear. The truncation to the lower 13 dimensions retains the spectral envelope and discards excitation frequency. Next, cepstral mean normalisation is performed for removing convolutional channel effects. Finally, if the norm of a vector of the 13 mean normalised cepstral coefficients (energy along with 12 cepstra representing the general spectral envelope) is higher than 1, then the vector is normalised in Euclidean space. This gives us the reduced variance of the search distance space. Corresponding confidence is taken as a proportion between best and worst relative distances from expectation in Euclidean space between corresponding audio logs. While having real-time computational efficiency, this approach requires much less RAM (15 MB versus 3 GB for fast cross correlation per 1 hour log).

3 Experimental Results and Visualisation

All results presented in this paper were achieved on a real life dataset of 10 social events (up to 1.5 hour each) with total 236 recorded subsegments (superposition of these events gives us 2360 possible combinations on the level of subsegments analysis), group of 8 socially connected people (who were using audio logging within the events) with personal audio-enabled palm devices (mobile phones and camcorders from 7 different manufactures). In figure 1 we illustrate how the length of the test subsegments influences the performance (the number of correctly clustered subsegments divided by the total number of test subsegments).

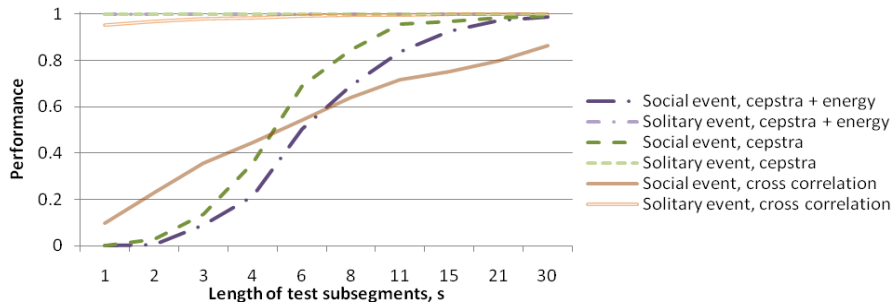

Fig. 1. Performance versus subsegment length of matching social context for the events with socially connected people and the events with no socially connected people involved.

We were used the fixed confidence threshold equals to 50% of subjective confidence, which is higher than the worst confidence for solitary events. This has the effect of minimising false detection of social events, though the application of dynamic threshold selection can further increase the total performance. The performance of shorter subsegments is lower due the real world variability of the data (noise, reverberation, non-stationarity, etc).

Fig. 2. Example of possible visualisation. Different socially connected people are automatically mapped on per day basis into personal memoirs of togetherness.

Figure 2 illustrates one of possible applications, targeting remote families (or any group of socially together people). When audio log is synchronised with an application the additional information about external media resources availability can be automatically retrieved via services-on-demand from social networks and mapped into the same personal memoirs, simplifying the navigation in tempo-social domain.

4 Conclusion

We have shown how the gap between user centric media devices and services-on-demand from social networks can be filled by automatic generation of tempo-social memoirs of togetherness. We found that the reliable memoirs can be generated from relatively short subsegments represented by small number of normalised cepstra. We have estimated that results surpass the performance of fast cross correlation, while requiring less resources. The achieved results give us the green light to further evaluate the presented concept from the privacy and the anxiety issues concerning being recorded everywhere and all the time, to concentrate on better understanding of the relations between psychoacoustic perception and social signal processing, to search for optimal ways of unobtrusive integration with existing applications.

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. I should like to extend my gratitude to Philip N. Garner, Herve Bourlard and John Dines for their valuable help and support at various stages of this work.

References

1. Wyatt, D., Choudhury, T., Kautz, H.: Capturing Spontaneous Conversation and Social Dynamics: A Privacy-Sensitive Data Collection Effort. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, USA (2007)
2. Farrahi, K., Gatica-Perez, D.: What Did You Do Today: Discovering Daily Routines from Large-Scale Mobile Data. Proceeding of the 16th ACM International Conference on Multimedia, Vancouver, Canada (2008)
3. Mermelstein, P.: Distance Measures for Speech Recognition, Psychological and Instrumental. In Pattern Recognition and Artificial Intelligence, pp. 374-388, C. H. Chen, Ed., Academic, New York (1976)
4. Ning, H., Roger, B. D., George T.: 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)
5. Choudhury, T., Pentland, A.: Sensing and Modeling Human Networks using the Sociometer. Proceedings of the 7th IEEE International Symposium on Wearable Computers, Washington, DC, USA (2003)
6. Korchagin, D.: Out-of-Scene AV Data Detection. Proceedings of the IADIS International Conference on Applied Computing, Rome, Italy (2009)