Challenges in Audio Processing of Terrorist-Related Data

Jodie Gauvain[1(✉)], Lori Lamel[2], Viet Bac Le[1], Julien Despres[1], Jean-Luc Gauvain[2], Abdel Messaoudi[1], Bianca Vieru[1], and Waad Ben Kheder[2]

[1] Vocapia Research, Orsay, France
{jodie,levb,despres,abdel,vieru}@vocapia.com
[2] CNRS-LIMSI, TLP, Orsay, France
{lamel,gauvain,benkheder}@limsi.fr
http://www.vocapia.com
http://www.limsi.fr/tlp

Abstract. Much information in multimedia data related to terrorist activity can be extracted from the audio content. Our work in ongoing projects aims to provide a complete description of the audio portion of multimedia documents. The information that can be extracted can be derived from diarization, classification of acoustic events, language and speaker segmentation and clustering, as well as automatic transcription of the speech portions. An important consideration is ensuring that the audio processing technologies are well suited to the types of data of interest to the law enforcement agencies. While language identification and speech recognition may be considered as 'mature technologies', our experience is that even state-of-the-art systems require customisation and enhancements to address the challenges of terrorist-related audio documents.

Keywords: Automatic speech recognition · Acoustic event detection
Language identification · Code switching

1 Introduction

This paper reports on recent research aiming to develop audio analysis technologies to facilitate access to information, helping investigators analysing terrorist-related activities to classify and search through audio or video documents. This research was conducted in the context of a European project focusing on multilingual multimedia data collected from the Web, potentially of interest in law enforcement investigations.

Analysis of this type of data poses a number of challenges rarely found in traditional broadcast data targeted by speech recognition systems. The challenges include: a wide range of recording environments with a variety of background

This work was partially financed by the Horizon 2020 project DANTE - Detecting and analysing terrorist-related online contents and financing activities and the French National Agency for Research as part of the SALSA project (Speech and Language technologies for Security Applications) under grant ANR-14-CE28-0021.

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11296, pp. 80–92, 2019.
https://doi.org/10.1007/978-3-030-05716-9_7

noises (heavy artillery, strong wind, rain, music, singing, crowd shouting, and other human or mechanically produced noises); the presence of many different native and non-native accents in multiple languages, language switching; and various speaking styles (preaching, chanting, shouting, whispering, ...).

So far, little work has looked into analysing such data, with investigations focusing more on telephone speech recordings for example. But today with the exponential amounts of audiovisual content posted daily on the Web, the growing threat of terrorism and the increasing use of Web platforms by terrorist organisations, it is essential to develnop solutions to efficiently process such content.

In addition to the well-known national and international terrorist investigation units, a growing number of international projects have started addressing the monitoring of such activities in multimedia data. LASIE (www.lasie-project.eu), RAMSES (www.ramses2020.eu), DANTE (www.h2020-dante.eu), PERICLES (www.project-pericles.eu), PROTON (www.projectproton.eu), TAKEDOWN (www.takedownproject.eu), TENSOR (www.tensor-project.eu), RED-ALERT (www.redalertproject.eu) and VICTORIA (www.victoria-project.eu) all aim at retrieving and processing multimedia contents linked to criminal activities for law enforcement purposes, but research is still in its early stages.

2 Audio Analysis Tasks

The proposed audio analysis solutions include identifying the language(s) of an audio document, transcribing speech into text and recognising specific acoustic events. Figure 1 gives a high level use of these technologies in the context of a tool to help humans analyse huge quantities of audiovisual data.

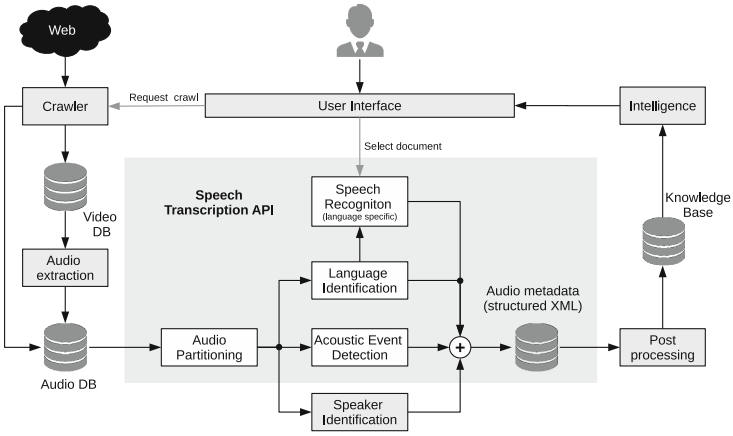

Fig. 1. Elements of the audio analysis process in a high-level context.

Automatic spoken language identification (LID) systems perform automatic detection of the spoken language(s), using the characteristics of the speech signal. LID can be used as a standalone technology, for instance for categorisation

purposes, or in association with other technologies, such as automatic speech recognition (ASR). ASR is used to automatically produce a transcript of what is said from the speech signal. Since ASR systems are generally language-specific, it is often useful to combine LID with speech recognisers to provide multilingual transcription functionality [1]. Finally, acoustic event detection (AED) is the task of automatically recognising different types of sounds (whether impulsive, continuous or intermittent) that can be of interest in an audio signal. AED can be used on its own, or in association with other technologies, bringing complementary information to automatic video analysis, for example. Speaker identification from audio is also shown in the figure but is not discussed in this paper.

The three tasks mentioned above all rely on an element called the audio partitioner. It is used to divide the acoustic signal into homogeneous segments, which are further combined into clusters. The partitioner uses a segmentation and labelling procedure based on an audio stream mixture model [2]. After detecting and eliminating non-speech segments, an iterative segmentation and clustering procedure is applied to the speech segments. Each resulting cluster represents roughly a speaker at a given acoustic condition (channel, background noise, etc.) and is assigned a unique label containing gender and channel information.

The data used in the testing phases for each of the audio analysis tasks is a corpus of unannotated terrorist propaganda videos retrieved from the Web. It contains roughly 500 h of audio and 400 h of speech, as detected by the audio partitioner. The tasks and technologies are described in the following sections, along with results and analyses.

3 Language Identification and Code-Switching Detection

State-of-the-art language recognition technology relies on statistical methods. The widely used phonotactic approach is based on the observation that phoneme sequences are distributed differently across languages [3]. The basic approach for LID was proposed in the early 90s, and relied on phone-based acoustic likelihoods [4]. The approach has since been extended to using parallel phone recognisers with phonotactic characteristics [6,7], lexical information [8] and phone lattices [9]. Recently, the i-vector framework, widely adopted in the speaker recognition community, has started being applied to the LID task [10,11].

Code switching (CS) happens when a speaker switches languages within or between utterances. The sociolinguistic implication and motivations for CS have been studied for several years [13–15]. CS is most commonly used by speakers exposed to some form of bilingualism, and generally in the context of spontaneous conversations. The presence of CS poses challenges both for LID and ASR.

Most LID research assumes that an audio document is in a single language, but depending on the task this is not necessarily true. Different approaches were explored to allow the LID system to analyse potential multilingual documents. One option is to determine the predominant language only and another is to output a list of most likely languages with associated scores. An alternative is to partition the audio into speech segments and detect the language of each segment. This approach is suitable for detecting relatively long language segments.

Indeed, LID is highly dependent on the segment duration, and performance can be significantly higher on long segments, for instance longer than 10 s. To ensure a minimal speech duration, LID can be applied to clusters of segments.

3.1 Experimental Conditions

Since LID is a classification problem based upon statistical models of speech, the models have to be trained on data that match the targeted data in order to achieve suitable accuracy levels. In this project, the targeted data consist of video documents containing propaganda or terrorist training instructions. As no task-specific training data were available, broadcast data (principally TV and radio news, talk shows, debates and interviews) were used. This type of data is easily available and was assumed to be the best match among the available corpora. The training data used in this work consists of 1295 h of broadcast news and broadcast conversation shows in 32 languages, collected during several R&D projects. The corpus contains speech from many speakers, several dialects and accents per language, and high variability in acoustic conditions. The amount of training data ranges from 11 to 142 h of speech/language. The test set is composed of the same types of data as used for training, with a total of 96 h of speech in 23 of the 32 languages, with at least 3 h/language.

The baseline LID system's phone decoders make use of HMM-GMM (Gaussian mixture models) acoustic models, whereas the improved system relies on phone decoders with output observation densities produced by (Deep) Neural Networks (DNN) [12]. These models were used to decode the language specific training data in order to estimate phonotactic constraints for each target language. For testing, each data sample is then processed by one or several phone recognisers. In addition, language-specific i-vectors were trained, and during test an i-vector is extracted for each segment and scored against each language-specific vector.

3.2 Experimental Results

The left side of Fig. 2 shows the language error rate (LER) of the baseline and improved LID systems as a function of the minimal cluster duration on a broadcast speech test set. As expected, the performance is seen to depend on the segment cluster duration, being higher for longer segments, and lower for shorter ones. The improved phonotactic system outperforms the baseline by up to 50% relative for the longer segments. The phonotactic and i-vector methods were compared using a single phone decoder. The i-vector system obtains better results on short speech segments (25.2% relative), whereas the phonotactic approach performs better on longer ones (14.3% relative).

Figure 2 (right) shows the LID output on the terrorist propaganda dataset using the segment-cluster mode for processing. Segments shorter than 10 s were removed in order to avoid sections falsely recognised as speech or mislabelled by the LID system. To ensure even higher accuracy, the language confidence scores provided by the system served to narrow down the segments retained. Of the

Duration (sec)	# clusters	System Baseline	System Improved	Relative Reduction
≥ 2.4	3798	6.0	4.3	29.2
≥ 4.0	3744	5.3	3.5	33.3
≥ 8.0	3638	4.4	2.7	38.9
≥ 12.0	3526	3.9	2.1	48.6
≥ 24.0	3060	2.2	1.1	50.7

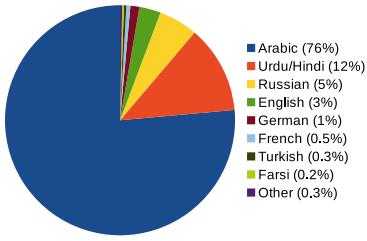

Fig. 2. Left: LER on a 23-language test set for the baseline and improved phonotactic LID systems as a function of cluster duration on an the internal broadcast data test set. Right: Proportion of speech detected per language in the terrorist propaganda videos.

files where speech was found, 16% were detected as containing more than one language, highlighting the presence of CS in this data.

4 Multilingual Speech Recognition

The last decade has witnessed major advances in speech and language technologies, which are becoming key components for analysing human communication in audio documents. The principles on which ASR systems are based on have been known for many years now, and include the application of information theory to speech recognition [16], the use of a spectral representation of the speech signal, of dynamic programming for decoding, and the use of context-dependent acoustic models [17]. Even though many of these techniques were proposed well over a decade ago, much of the recent progress is due to the availability of large speech and text corpora, and improved processing power which have allowed more complex models and algorithms to be implemented.

Transcription performance varies substantially across data types. While for well-trained ASR systems word error rates (WER) can be in the range of 5–10% on carefully prepared speech, the error rate is easily doubled or tripled for spontaneous speech or in degraded acoustic conditions (WER above 50%). It is widely acknowledged that the performance of a speech recogniser is strongly dependent upon the task, which in turn is linked to the type of user, speaking style, environmental conditions, etc. The emergence of new online-terrorist communities being so recent, very little work has been done on processing the audio contents that they generate. The main challenges lie in the intrinsic variety of such recordings, which can range over any type of quality and real-life situation, not to mention diversity of speakers, emotions, accents, languages, and in particular the use of multiple languages (Code-Switching). All these conditions require specific research in order to go beyond the state-of-the-art.

4.1 ASR System Overview

Most ASR systems have five main components: an audio partitioner, an acoustic model, a statistical language model (LM), a pronunciation dictionary, and a

word recognizer [2]. As for LID, no terrorist-related data was available for system development, so broadcast data was used for training and testing as it was assumed to be the best match. The audio partitioner, designed for broadcast speech, generates a sequence of non-overlapping segments and groups them into clusters. Acoustic and language models were trained using statistical methods on large quantities of data. The language model training data includes manual transcriptions of recordings, written dialogues, news and other types of sources that can be gathered from the Web. The acoustic models are triphone-based Hidden Markov Models, with output observation probabilities given by DNNs [12]. For each language, the acoustic models were built using state-of-the-art discriminative training methods and trained on several hundred hours of annotated data (audio recordings and their associated transcriptions). The phone sets cover the language-specific phones and special units to model silence, breath and filler words. The pronunciation dictionaries are built with grapheme-to-phoneme rules derived from linguistic knowledge, complemented with exception rules as needed.

4.2 Experimental Conditions and Results

Six languages are targeted for ASR: Arabic, English, French, Italian, Portuguese and Spanish. Table 1 displays the WER for the baseline and enhanced systems on internal broadcast datasets. Developments for the improved systems involved, in particular, new acoustic modeling based methods on time-delay neural networks (TDNNs) [18] and acoustic data augmentation, including speed and volume perturbation, addition of background noise and reverberation. Both methods have been proven effective to make models more robust to noisy environments and help cope with mismatches between training and testing data [19,20].

Table 1. Word error rate (WER (%)) on an internal broadcast speech development data set with at least 3 h of data from a minimum of 20 speakers per language.

Language	Arabic	English	French	Italian	Portuguese	Spanish
Baseline	9.6	13.3	13.7	9.9	14.2	11.2
Improved	8.9	10.5	11.3	8.1	13.8	10.2

Figure 3 shows an audio excerpt with segments in 3 different languages (English, Arabic and French), with their corresponding automatic transcripts. LID and ASR were jointly applied to produce multilingual speech-to-text.

A subset (7 h) of the terrorist propaganda dataset was manually selected and annotated for testing. This was performed in Arabic, because of its overall predominance in the data, and in English, for demonstration. The ASR system was used to produce a transcript of the audio, and the output was scored against the reference transcripts. Initial results show a considerable decrease in accuracy when compared to those reported in Table 1. The baseline Arabic system obtains a WER close to 30% on this data (compared to 9.6 on the broadcast speech

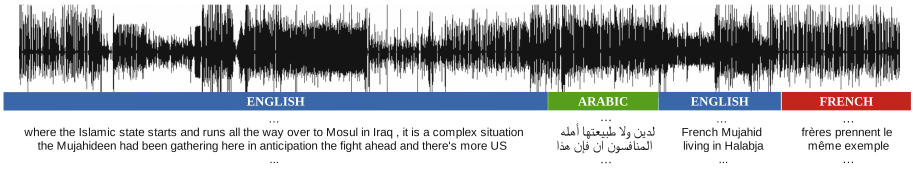

Fig. 3. ASR an audio excerpt containing segments in 3 different languages.

test set), and the English system nearly reaches 44% (compared to 13.3% on a broadcast speech test). The enhanced English system was also tested on this dataset, resulting in a 39% WER (11.5% relative gain compared to the baseline).

These results, which can in part be attributed to the difference between the target and and training data, highlight the difficulty of the task at hand.

Some challenging aspects of the terrorist propaganda audio were noticed across all languages. The speech partitioning is perturbed by the strong presence of chanting and preaching that can easily be mistaken for speech, as well as a wide range of background noises. Many files have a relatively low audio quality: the microphones are often placed far from the speakers, sometimes bringing environmental noise to the foreground instead of the speech. In addition, the ASR language models and lexicons are not adapted to this data, the vocabulary and formulations being quite different from those of the training data. For the English speech in particular, strong accents of non-native speakers are sources of many errors. The speech also contains many hesitations and grammatical errors that do not match well with the language models. CS with Arabic is also omnipresent, and over 8% of the words in the reference transcripts are not in the ASR lexicon. As shown in Table 2, even words of Arabic origin that are now commonly used in English are often missed as their Arabic pronunciations can be very different from those in the English lexicon.

Table 2. Common Arabic words with their English and Arabic pronunciations (mapped to the English phone set). Differences are shown in color.

Word	Occurences	Recognized	English pron	Arabic pron
Jihad	36	24	JIhad JIh@d Jihad Jih@d	Zihad
Allah	169	3	@lx	alah
Mujahideen	24	11	myuZxhxdin muZxhxdin myuJxhxdin muJxhxdin	muZahidin

5 Acoustic Event Detection

Sounds carry a large amount of information about our environment and the physical events that take place in it. Humans naturally perceive the sound scene around them (busy street, office, etc.), and can recognise individual sound sources (car passing by, footsteps, etc.). For decades researchers have been fascinated with the idea of machines that could hear and understand audio content just like humans do, referred to as 'machine listening'. Developing signal processing methods to automatically extract this information has huge potential in several applications. The goal of AED is to label temporal regions within an audio recording, determining the start, end and the nature of sound instances. The output can be exploited jointly with other technologies, such as image or video recognition services, bringing valuable complementary information to the table.

Interest in AED has been increasing over recent years, with public challenges, such as DCASE (http://dcase.community), helping to boost research in the field. Unfortunately, most benchmarks and available datasets are not very relevant to this project. In addition, published detection and classification performances on similar tasks are still quite low. Even when only trying to detect a few categories of events, performance remains relatively low (compared to what is seen nowadays with ASR). This illustrates the difficulty of the task and the progress that is still to be made in order to reliably recognise sounds in realistic soundscapes, where multiple sounds are present, often simultaneously, and distorted by the environment. It is important to note that most challenges primarily address classification of events, eliminating altogether the detection stage which adds another level of difficulty. The closest work to ours is that of Google AudioSet described in [21].

5.1 System Description

Neural Networks have proven to be very efficient for speech processing activities, recently resulting in a significant leap in system accuracy. But for AED, the trend to use Deep Convolutional Neural Networks [22] has shown less convincing results in the 2017 DCASE challenge (http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/index) [23]. This is probably due, in part, to the lack of well-annotated data to work with, and of course, to the inherent complexity of the task. In order to incorporate the AED system as shown in Fig. 1, implementing CNN-based models would have required major modifications to the structure of the partitioner. Therefore, for initial experiments, the same acoustic feature extraction methods as for speech were used, allowing a simpler incorporation of new events into the existing technology.

Sounds of interest were selected in collaboration with law enforcement partners, and further narrowed down according to their availability in publicly available datasets. Out of 15 corpora inventoried, of various sizes and containing many sub-corpora, only a few covered the audio events of interest, and Google AudioSet (https://research.google.com/audioset) was the only one to cover all of them. It contains over 2 million semi-automatically labeled 10-second sound

clips drawn from YouTube videos with a hierarchical ontology [21], partially validated by humans.

Given the large disparities in the available data (both in quantity and quality), only a few events of critical importance were experimented with in a first validation stage. In addition to speech, four acoustic events were focused on: explosions, shootings (gunshots and machine guns), and Nasheed (singing). The Nasheed is a work of vocal music that usually makes reference to Islamic beliefs, and is meant to inspire Muslims to practice Jihad. This type of singing, shown on the left of Fig. 4, having formant structure similar to speech, is often present in terrorist propaganda recordings and was designated as an important sound to detect by the law enforcement agencies.

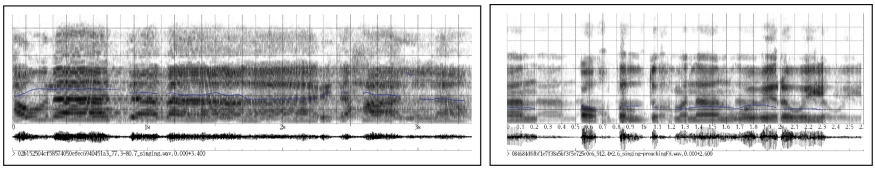

Fig. 4. Spectrograms of singing with background music (left) and preaching (right).

In order to be seamlessly integrated in the audio partitioner, GMMs were used to model the acoustic events. The GMMs for the audio segmentation and labelling procedure use basically the same acoustic feature vector as what is typically used for ASR with the exception that it does not include the energy, but does use the delta energy parameters. For speech and general music, we used the MUSAN music, speech and noise corpus [24], composed of 109 h of precisely annotated audio. GMMs for the other acoustic events were trained using data extracted from the Google Audioset corpus. Finally, the model for the Nasheed was trained on manually annotated data from real terrorist propaganda videos, since it was not included in the AudioSet ontology.

5.2 Experimental Results

Table 3 gives the results of a manual validation of a randomly selected subset of the acoustic events detected in 100 h of audio from the propaganda corpus.

It can be seen that the number of correct detections (validated) is highest for singing, for which carefully annotated training segments were used. There are more false alarms than correct detections for the other 3 categories, with the largest number of false alarms on gunshots. One explanation may be that since these models were trained on 10-second AudioSet samples, there can be other sounds in the segment, which may impact shorter events more than longer ones.

Figure 5 shows a spectrogram of an audio segment classified as machine gunfire. While the regular burst seen in the signal and the spectrogram correspond to machine gunfire, the three darker, longer bursts are explosions overlapping with

Table 3. Manual validation of a random subset of approximately 400 automatically detected acoustic events. The numbers correspond to the total number of detections, the correct/incorrect/unclear detections (Validated/False Alarms).

Acoustic event	Detected	Validated	False alarm	Unclear
Explosion	1564	95	149	63
Gunshot	1207	36	402	22
Machine gun	927	71	293	27
Singing	4745	154	143	11

the gunfire. There are many other polyphony instances, with impulsive acoustic events overlapping other continuous or repetitive events such as singing, wind, speech, steps, etc. Many of these sounds are difficult even for humans to distinguish, for example machine gunfire can sound like fireworks or a loud engine.

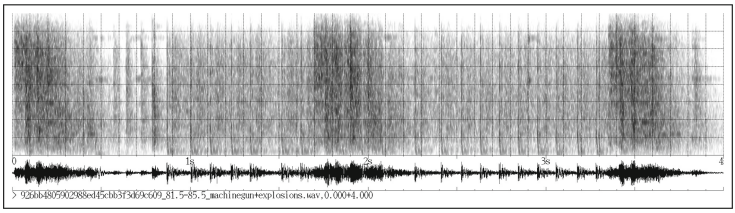

Fig. 5. Spectrogram illustrating explosions alternating with machine gunfire.

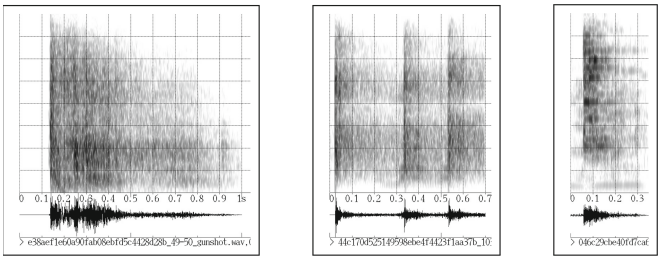

Fig. 6. Spectrograms of single and multiple gunshots (left, middle) and a clang (right).

5.3 Challenges

Some major challenges are still to be tackled. First and foremost, data annotation needs to be improved. The study of the different corpora on AED, and

the manual annotation process that was begun show that, even for humans, annotating sounds is a very difficult task. When one is actually within an environment, many sources of information are received about what is happening, but when listening to audio a posteriori without knowing the context, it is difficult for the human ear to distinguish between similar sounds (as shown by the last column in Table 3). This is also illustrated in Fig. 6 which shows spectrograms of a single and multiple gunshots (left, middle) and a clang that was mistakenly detected as a gunshot. Many acoustic events can easily be confused, and have almost identical spectral features. For example a 'bang' can be an explosion, a gunshot, thunder, a firecracker, etc. A first step in future works will need to be the careful selection of events to annotate, the definition of exactly how they should be annotated and the use of as much context as possible to annotate them (exploiting the video images, for example, when available).

A second major difficulty for AED tasks is the issue of polyphony. Unlike speakers who usually try to take turns speaking, there can be an infinite number of overlapping sound events and it is therefore nearly impossible to try to detect and recognize them all. The majority of work on AED treats the sound as monophonic, assuming that only one event is detectable at a time, but, in most real-world situations, sounds overlap and events of interest can co-occur (as shown in Fig. 5 which has explosions overlapping with machine gunfire).

6 Summary and Discussion

This paper has presented some of the challenges in the automatic processing of terrorist-related audio data found on the Web and some of the initial progress made in addressing these challenges. Concerning language identification and code switching, phonotactic and i-vector methods have been explored, and improved decoders developed. Segment-cluster based LID was introduced to handle multiple languages in an audio document.

Concerning speech recognition, improved acoustic models have been developed for the 6 languages of interest in the project using the latest acoustic modelling techniques. Acoustic data augmentation was used to increase the amount and variability of the training data thereby improving genericity and reducing the mismatch between the training and test data. Future developments will address improving the language model components by locating texts that are close to the targeted data, and improving the pronunciation lexicon for accented speech. We have also started exploring bilingual decoding as a means of handling code-switching, where two ASR systems process the data in parallel, allowing a language switch at each word.

It is interesting to note that it was considered a huge challenge when the National Institute of Standards and Technology (https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation) first proposed the task of automatically transcribing broadcast news data back in the 90's. Until that time ASR had mainly addressed processing of read speech, dictation or simple constrained tasks that did not need to deal with heterogeneous data, multiplicity of speakers and acoustic conditions, speech in the presence of music, etc. Today the transcription of

broadcast news data is considered a relatively simple task compared to less formal data types such as conversational speech, amateur youtube videos and multiparty meetings [25]. Therefore, we can hope that similar progress will be made in the future at transcribing challenging terrorist-related audio.

Acoustic event detection is still in its early stages, but research on comparable problems such as object detection in images has recently shown astonishing results. The polyphony issue is still a long way from being solved, and is one of the reasons why AED is considered by many specialists as a very difficult task. However, with a better annotation process, machine performance is expected to improve. A semi-supervised method relying both on automatic recognition and human validation at a finer scale than was used for the AudioSet labels could be the key.

References

1. Vu, N.T. et al.: A first speech recognition system for Mandarin-English code-switch conversational speech. In: IEEE ICASSP (2012)
2. Gauvain, J.L., Lamel, L., Adda, G.: Audio partitioningt and transcription for broadcast data indexation. Multimed. Tools Appl. **14**, 187–200 (2001)
3. House, A.S., Neuburg, E.P.: Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. JASA **62**(3), 708–713 (1977)
4. Gauvain, J.L., Lamel, L.: Identification of non-linguistic speech features. In: Human Language Technology (HLT 1993), pp. 96–101. ACL (1993)
5. Lamel, L., Gauvain, J.L.: A phone-based approach to non-linguistic speech feature identification. Comput. Speech Lang. **9**(1), 87–103 (1995). https://doi.org/10.1006/csla.1995.0005
6. Zissman, M.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio **4**, 31–44 (1996)
7. Benzeghiba, M. Gauvain, J.L., Lamel, L.: Improved n-gram phonotactic models for language recognition. In: Interspeech (2010)
8. Kadambe, S., Hieronymus, J.: Language identification with phonological and lexical models. In: IEEE ICASSP (1995)
9. Gauvain, J.L., Messaoudi, A., Schwenk, H.: Language recognition using phone lattices. In: ICSLP, pp. 1283–1286, Jeju Island (2004)
10. Dehak, N. et al.: Language recognition via i-vectors and dimensionality reduction. In: Interspeech, pp. 857–860, Florence (2011)
11. Martinez, D. et al.: Language recognition in iVectors space. In: Interspeech (2011)
12. Hinton, G., et al.: Deep neural networks foracoustic modeling in speech recognition. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
13. Weinreich, U.: Languages in Contact. Mouton, The Hague (1953)
14. Demby, G.: How code-switching explains the world (2013)
15. Amazouz, D., Adda-Decker, M, Lamel, L.: Addressing code-switching in French/Algerian Arabic speech. In: Proceedings of Interspeech 2017, pp. 62–66 (2017)
16. Jelinek, F.: Continuous speech recognition by statistical methods. Proc. IEEE **64**, 532–556 (1976)

17. Schwartz, R. et al.: Improved hidden Markov modeling of phonemes for continuous speech recognition. In: IEEE ICASSP, vol. 3, pp. 35.6.1–35.6.4 (1984)
18. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Interspeech (2015)
19. Cui, X., Goel, V., Kingsbury, B.: Data augmentation for deep neural network acoustic modelling. In: IEEE ICASSP, pp. 5619–5623 (2014)
20. Ragni, A., et al.: Data augmentation for low resource languages. In: Interspeech, pp. 810–814, Singapore (2014)
21. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: IEEE ICASSP, pp. 776–780 (2017)
22. Hershey, S. et al.: CNN architectures for large-scale audio classification. In: IEEE ICASSP, pp. 131–135 (2017)
23. Takahashi, N. et al.: Deep convolutional neural networks and data augmentation for acoustic event detection, arXiv preprint arXiv:1604.07160 (2016)
24. Snyder, D., Chen, G., Povey, D.: MUSAN: a music, speech, and noise corpus, CoRR abs/1510.08484 (2015). http://arxiv.org/pdf/1510.08484v1.pdf
25. Martin, A. Garofolo, J.: NIST speech processing evaluations: LVCSR, speaker recognition, language recognition. In: IEEE Workshop on Signal Processing Applications for Public Security and Forensics, pp. 1–7 (2007)