"Sheldon speaking, bonjour !" – Leveraging Multilingual Tracks for (Weakly) Supervised Speaker Identification

Hervé Bredin
Anindya Roy
Nicolas Pécheux
Alexandre Allauzen
–
LIMSI - CNRS
BP 133, Orsay, France
bredin@limsi.fr

ABSTRACT

We address the problem of speaker identification in multimedia data, and TV series in particular. While speaker identification is traditionally a supervised machine-learning task, our first contribution is to significantly reduce the need for costly preliminary manual annotations through the use of automatically aligned (and potentially noisy) fan-generated transcripts and subtitles. We show that both speech activity detection and speech turn identification modules trained in this weakly supervised manner achieve similar performance as their fully supervised counterparts (*i.e.* relying on fine manual speech/non-speech/speaker annotation). Our second contribution relates to the use of multilingual audio tracks usually available with this kind of content to significantly improve the overall speaker identification performance. Reproducible experiments (including dataset, manual annotations and source code) performed on the first six episodes of *The Big Bang Theory* TV series show that combining the French audio track (containing dubbed actor voices) with the English one (with the original actor voices) improves the overall English speaker identification performance by 5% absolute and up to 70% relative on the five main characters.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms

Keywords

speech activity detection; speaker identification; multimedia data; weak supervision; multilingual fusion

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM'14, November 03 – 07 2014, Orlando, FL, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654929.

1. INTRODUCTION

Rich in interactions between characters and enjoying a wide fan base, movies or TV series such as *Harry Potter* or *The Big Bang Theory* are a potential source of data for both natural language processing applications (*e.g.* summarization) and information retrieval tasks (*e.g.* fans may like to retrieve all scenes where *Leonard* invites *Penny* to dinner). Second screen applications could rely on the textual transcription of dialogues between the main characters of a TV series to provide an efficient way to browse a TV series scene by scene [6] or from punchline to punchline (*e.g. Sheldon*'s famous *"Penny, Penny, Penny"* or *"That's my spot!"*) [18].

The main objective of this work is to automatically augment TV series with additional metadata that might eventually lead to those novel multimedia retrieval applications. While automatic speech transcription (or DVD subtitles when they are available) can be used to gain a clear insight of what is being said at any time, the metadata describing which character pronounced a particular line is usually missing, though this information is crucial for content-based video retrieval purposes. We propose to apply speaker identification to automatically obtain this missing information. Figure 1 summarizes our contributions towards this objective.

Thanks to speaker recognition evaluations (SRE) organized by NIST since 1996 [24] and more recent initiatives such as ESTER, ETAPE and REPERE [21], many studies have been devoted to speaker recognition in conversational phone calls or radio and TV broadcast news. However, to the best of our knowledge, it is the first time speaker recognition is also applied to TV series. Section 2 details the different challenges raised by this new task and introduces our baseline approach to solve the problem.

The main limitation of automatic speaker identification is that it relies on a time consuming manual annotation step for the initial training of speaker models. To address this issue in a fully unsupervised manner, a typical approach is to detect speaker names from speech transcripts or subtitles and try to propagate them to speaker clusters [10, 39, 26]. However, reported results show that unsupervised speaker identification performance decreases rapidly (down to less than 30% accuracy) when relying on automatic modules (*i.e.* speech transcription, name detection and speech turns clustering) [16, 20].

Our first major contribution – described in Section 3 – is to show how one can dramatically reduce the cost of

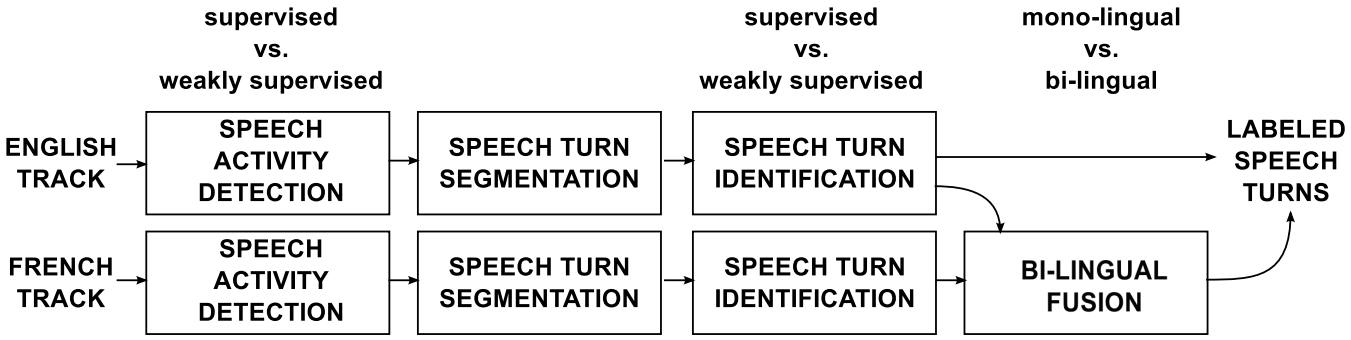

Figure 1: Speaker identification in TV series. Our contributions include weakly supervised speech activity detection, weakly supervised speaker identification, and bi-lingual speaker identification.

this preliminary manual annotation step, with almost no performance loss, via automatic temporal alignment of fan-generated transcripts with corresponding TV series episodes. Alignment of screenplays and transcripts has been used before in the computer vision community for face recognition in movies or TV series in [17, 34, 13, 36, 4]. Though these annotations are often noisy and incomplete, experiments show that the overall speaker identification performance is only slightly degraded, in comparison with the same approach based on precise – but costly – manual annotation of the same training set.

Our second main contribution relates to the use of multilingual audio tracks to further improve speaker identification. While the combination of multiple sources of information (such as voice, face and text) has been used before for multimodal person identification [8, 30], we are not aware of any previous work combining synchronized audio tracks in multiple languages to improve character (or more generally, speaker) identification as done in this work.

As a matter of fact, with the advent of digital broadcasting, most movies and TV series episodes are distributed at least in both the production original language and the viewers' language (with dubbed voices). Section 4 describes how one can take advantage of these multiple audio tracks to reach nearly perfect speaker identification among main characters.

The experimental protocol described in Section 5 follows the reproducible research principle: not only is the used corpus reproducible locally, but the necessary source code to reproduce the experiments is also distributed on the corpus webpage (`tvd.niderb.fr`). Finally, results are reported and discussed in Section 6. Section 7 concludes the paper.

2. SPEAKER IDENTIFICATION

In this section, we describe the speaker identification approach that will serve as a baseline for the rest of the paper. The upper part of Figure 1 summarizes it graphically. As this is the first time speaker identification in TV series is addressed, we first motivate several design choices by a few characterics specific to TV series.

Spontaneous speech While speech is usually prepared in TV broadcast news, dialogues between characters can be considered as spontaneous speech (even though they are acted); thus resulting into short speech turns with lots of fast interactions between characters.

Small set of characters While a large number of speakers are supposed to be recognized in above-mentioned evaluation campaigns, the number of characters in TV series is very limited and mostly main characters are of interest to the end user; the task can therefore be considered as closed-set speaker identification with a small number of targets.

Clean audio The audio track of a TV series is usually the result of a well-controlled production pipeline where the speech signal is carefully recorded on a dedicated channel before being mixed with other audio channels (*e.g.* music, stage noise or recorded laughs). Even though this *clean speech channel* is not available directly, the speech signal resulting from controlled post-production is still much cleaner than if it was recorded in a noisy environment.

2.1 Speech activity detection

Speech activity detection is addressed as a two-classes classification problem. Given the input audio stream, one has to decide for each instant whether a character is currently speaking (*speech* class) or not (*non-speech* class).

While standard speech activity detection usually divides the *speech* class into several sub-classes (*e.g.* clean speech, noisy speech or speech over music) [3], we do not. Indeed, as stated earlier, TV series audio tracks result from a careful post-production process leading to much higher audio quality than phone calls or broadcast news historically addressed by the community.

A two-states hidden Markov model (HMM) is used to tackle this problem: *speech vs. non-speech*. The states emission probability distributions are modeled by Gaussian mixture models. Given annotated (either manually or automatically) audio tracks and acoustic features extracted from them, the HMM parameters are estimated using an iterative Expectation-Maximization (EM) approach: the Baum-Welch algorithm [31].

At test time, given a previously unseen audio track, the Viterbi algorithm [31] is used to estimate the optimal sequence of hidden states (*i.e. speech* or *non-speech*). Finally, a median filter with a sliding window of 250 ms is applied to this sequence to get rid of too short *speech* (or *non-speech*) segments.

2.2 Speech turn segmentation

Due to the rapid interactions between characters in TV series, it is frequent that one continuous speech segment contains speech turns from multiple interacting characters. Therefore, before trying to label speech segments, they are further segmented into smaller homogeneous segments by detecting speaker changes [12]. This is achieved by detecting every maximum of the local Gaussian divergence $G(w_L, w_R)$ between two adjacent sliding windows w_L (left) and w_R (right) of 1 second. The Gaussian divergence is defined as follows:

$$G(w_L, w_R) = (\mu_R - \mu_L)^T \cdot \Sigma_L^{-1/2} \cdot \Sigma_R^{-1/2} \cdot (\mu_R - \mu_L) \quad (1)$$

where the set of acoustic features of each window is modeled as a Gaussian $\mathcal{N}(\mu, \Sigma)$ with diagonal covariance matrix Σ [3].

2.3 Speech turn identification

In this paper, we are only focusing on recognizing main characters as they are usually the ones of interest for the user willing to retrieve a particular moment in a TV series. Moreover, as shown later in Section 6, secondary characters only account for 10% of speech time in a TV series such as *The Big Bang Theory*. Therefore, we address the issue of identifying speech turns as a closed-set speaker identification problem: for each speech turn t, decide which of the N main characters is currently speaking.

We rely on a standard Gaussian mixture model (GMM) system based on adapted universal background model (UBM). It has proved to be very successful for text-independent speaker recognition, since it allows for robust estimation of speaker models λ_i even with a limited amount of enrollment data [32].

Given audio segments annotated as speech turns and acoustic features extracted from them, the UBM parameters are estimated using the iterative Expectation-Maximization (EM) algorithm. Then, for each main character i, a speaker-specific GMM λ_i is trained by MAP adaptation [19] of the means of the UBM.

While standard speaker identification algorithms (for telephone calls or broadcast news) usually rely on a preliminary step of feature normalization to compensate for channel and speaker variatibility [2], we do not. Indeed, as already stated before in Section 2.1, the audio post-production pipeline in TV series results in a readily normalized audio channel.

At test time, given a speech turn t and a target identity i, the speaker identification score ρ_{ti} is defined as the standard log-likelihood ratio [32]. Finally, the decided identity i^* of speech turn t is obtained as the one with the largest score: $i^* = \text{argmax}_{\ i}\, \rho_{ti}$.

A common practice for speaker identification in broadcast news is to rely on a preliminary speaker diarization step (*i.e.* speech turns clustering) and perform identification at cluster level rather than speech turn level [38]. However, because of the specificity of dialogues in TV series (spontaneous speech, short speech turns and fast interactions between characters), we found that state-of-the-art speaker diarization approaches do not perform well for this kind of content and tend to group speech turns of multiple interacting speakers into a unique cluster: we chose to perform identification at speech turn level, rather than at (likely noisy) cluster level.

3. REDUCING MANUAL SUPERVISION

Both speech activity detection and speech turn identification steps introduced in the previous paragraphs are supervised machine-learning techniques in need of on annotated training data to learn speech, non-speech and speaker models. However, as already stated, this mandatory annotated data is usually very costly (both in terms of time and human resources) to produce manually. Furthermore, while speech/non-speech classification models are quite generic and can be trained once and for all, speaker models are specific to a TV series and need to be updated for every new appearing character.

Named speaker identification techniques [10, 20, 7] could be used to achieve fully unsupervised speaker identification. However, this type of approaches rely on the combination of several error-prone processing stages (including automatic speech transcription, named entity detection, speaker diarization and name propagation), leading to overall bad performance.

Luckily, fan-generated manual transcripts containing speaker identities can be downloaded from the Internet for a limited number of episodes. Once temporally aligned with subtitles, they constitute a reliable source of annotations that can be used to train speaker models, consequently applied to the remaining episodes.

3.1 Aligning transcripts with subtitles

As shown in Figure 2, subtitles (a) typically consist of a sequence of dialogue lines $\mathcal{S} = \{s_i\}_{1 \leq i \leq N}$ associated with a time span. However, they do not provide information about speaker identity. On the other side, manual transcripts (b) consist of a sequence of dialogue lines $\mathcal{T} = \{t_j\}_{1 \leq j \leq M}$ each associated with a speaker name, but no temporal information. The objective of automatic alignment is to merge time spans from subtitles with speaker identities from manual transcripts – thus leading to the availability of speaker time spans (e).

3.1.1 Pre-processing

Subtitles may originally cover multiple dialogue lines from several speakers. In such cases, the subtitle time span is divided and allotted to each line proportionally to their number of words. For instance, the first subtitle of Figure 2 (a) is divided into two dialogue lines s_1 (*"...it will not've gone through both slits"*) and s_2 (*"Agreed."*). This simple heuristic is definitely not optimal as it may cut those multi-speaker subtitles at incorrect positions. More advanced techniques could be used instead (*e.g.* forced alignment of text and audio).

3.1.2 Word overlap statistics

In the case of *The Big Bang Theory* TV series, standard word-level alignment [17, 35, 27] in English reveals that around 97% words in manual transcripts have an exact match in subtitles while around 98% words in the subtitles have a match in manual transcripts. The mismatched cases are mostly due to deliberate dropping of uninformative words, simplification of sentences or contractions in the subtitles, *e.g. "it will not have gone"* in manual transcript t_1 is contracted to *"it will not've gone"* in subtitle s_1.

However, in the French language, the situation is less advantageous. While French subtitles can be obtained from DVDs, manual transcripts are not directly available in the

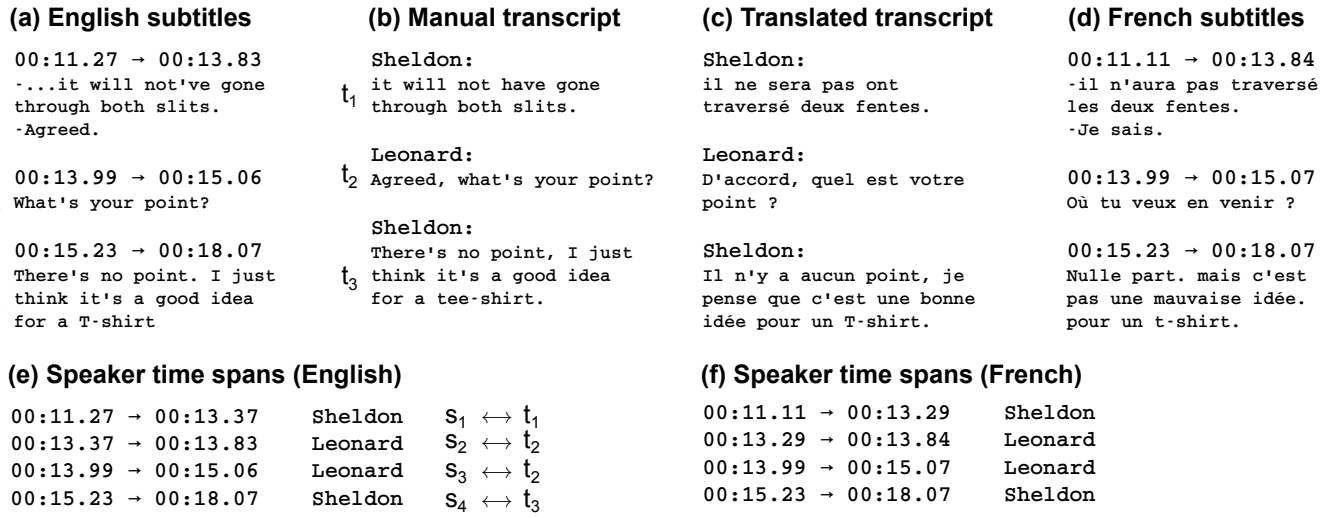

Figure 2: DVD subtitles (a and d), available English manual transcripts (b), English manual transcripts automatically translated into French (c) and time spans automatically annotated with speaker identity obtained after alignment of subtitles with transcripts (e and f).

French language and an additional step of automatic translation (described later in Section 4.2) is needed to obtain a French version of the English manual transcripts. A similar analysis of word-level alignment in French reveals that only 26% words in the translated manual transcripts are matched with the same word in French subtitles, while 32% words in French subtitles have a match with the same word in translated manual transcripts.

3.1.3 Alignment algorithm

Based on this observation, we chose not to use the standard word-level alignment algorithm as in [17, 35]. Instead, we consider manual transcripts and subtitles as *comparable corpora* and use a dedicated sentence-level alignment algorithm [28]. Let $\mathcal{W}$ be the overall set of words used either in manual transcripts or subtitles. Each subtitle or transcript line $\ell \in \mathcal{S} \cup \mathcal{T}$ is then described by a TF-IDF vector φ_ℓ defined as follows:

$$\forall w \in \mathcal{W}, \quad \varphi_\ell(w) = \mathrm{TF}_\ell(w) \cdot \log\left(\frac{N+M}{\mathrm{DF}(w)}\right) \tag{2}$$

where $\mathrm{TF}_\ell(w)$ is a binary indicator of whether word w occurs in (subtitle or transcript) line ℓ, N is the number of subtitle lines, M is the number of transcript lines and $\mathrm{DF}(w)$ is the total number of lines in which word w occurs. The local similarity between a subtitle s and a manual transcript t is defined as the cosine similarity between their respective TF-IDF vectors:

$$\sigma(s,t) = \frac{\sum_{w \in \mathcal{W}} \varphi_s(w) \cdot \varphi_t(w)}{\sqrt{\sum_{w \in \mathcal{W}} \varphi_s^2(w) \cdot \sum_{w \in \mathcal{W}} \varphi_t^2(w)}} \tag{3}$$

The standard Dynamic Time Warping (DTW) approach is then applied to obtain the alignment. First, the global alignment score $\kappa(s_i, t_j)$ is computed as follows:

$$\kappa(s_1,t_1) = \sigma(s_1,t_1) \tag{4}$$

$$\kappa(s_i,t_j) = \sigma(s_i,t_j) + \max\begin{cases}\kappa(s_{i-1},t_{j-1})\\ \kappa(s_{i-1},t_j)\\ \kappa(s_i,t_{j-1})\end{cases} \tag{5}$$

Then, we find the best alignment path by backtracking from $\kappa(s_N, t_M)$ to $\kappa(s_1, t_1)$. Pairs of lines (s_i, t_j) on this path are considered as aligned and time spans from the subtitle lines are merged with speaker names from the aligned manual transcript lines. This results in a sequence of time spans annotated with speaker identity as shown in columns (e) and (f) of Figure 2.

3.2 Weak supervision

This automatic alignment process removes the need for costly manual annotations and allow us to achieve weakly supervised speaker identification:

Speech activity detection. Subtitles time spans are marked as *speech* segments, while the rest of the audio track is marked as *non-speech*. Then, the HMM-based approach is trained the usual way as described in Section 2.1.

Speech turn segmentation is left unchanged as it does not rely on any supervision in the first place.

Speech turn identification. Automatically labeled subtitles are used as groundtruth for training of the identification module described in Section 2.3.

However, those automatic annotations do not have the same quality as manual annotations. Figure 3 illustrates this issue for both speech activity detection and speech turn identification. In particular, by comparing time spans of speech turns and DVD subtitles, one notices that DVD subtitles tend to last longer than the corresponding speech turns and therefore overlap non-speech regions (colored as light grey

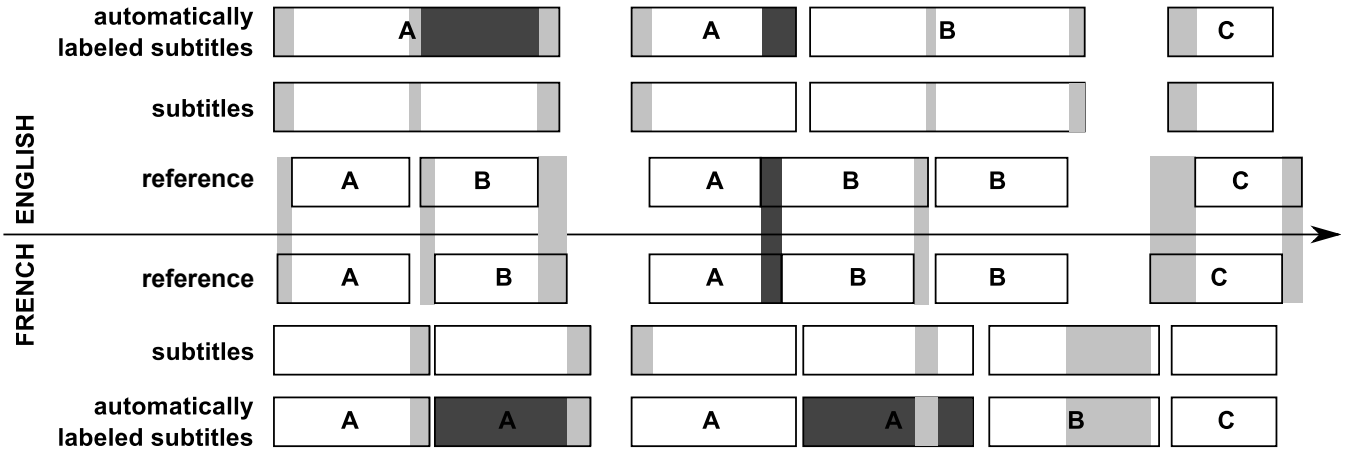

Figure 3: Badly segmented (light grey) regions in subtitles highlight the fact that subtitles time spans only provide coarse annotations for weakly supervised training of speech activity detection. Similarly, incorrectly labeled (dark grey) regions in automatically labeled subtitles highlight that they provide noisy supervision for speaker identification training. Finally, comparing English and French reference speech turns shows that they are approximately synchronized and therefore can be combined for bilingual speaker identification.

regions in Figure 3). This is actually deliberate because subtitles need to be displayed long enough for the viewer to be able to read them entirely. More precisely, statistics reported in column *subtitles* of Table 1 indicate that subtitles do cover most (95.6%) of *speech* regions. However, as anticipated, they also contain 25.4% of *non-speech* regions.

Similarly, errors in the automatic alignment of manual transcripts and subtitles may results in noisy speaker labels (colored as dark grey regions in Figure 3). However, we will show in Section 6 that the performance of both speech activity detection and speech turn identification is not degraded when those noisy annotations are used in place of manual annotations.

4. BILINGUAL IDENTIFICATION

In an increasingly multilingual and multicultural world, many TV programs come as rich multi-lingual content. TV series or movie DVDs, for example, usually contain multilingual audio tracks and/or subtitles. Original characters' voices are dubbed in alternative languages by professional voice actors. In a given language, each character is always dubbed by the same voice actor and no two main characters share the same voice actor.

While a speaker identification system may have trouble distinguishing the voices of two characters in the original language, it is quite unlikely to be also the case with their dubbed voices (and reciprocally). Hence, we show how one can take advantage of these complementary sources of information.

4.1 Bilingual fusion

In this work, we focus on a bilingual (English + French) fusion approach to make the most of the multi-lingual additional audio tracks and improve speaker identification in the original language (English). As illustrated in Figure 3, and for obvious lip-sync reasons, the French speech turns tend to strongly follow the segmentation in the original English language, though they are not perfectly synchronous. Therefore, we propose to train two distinct mono-lingual speaker identification systems (one for English and one for French) and combine them at score level:

$$\rho_{ti} = \alpha \cdot \rho_{ti}^{\mathrm{EN}} + (1-\alpha) \cdot \rho_{ti}^{\mathrm{FR}} \quad (6)$$

where $\alpha \in [0,1]$ is a weighting coefficient.

While we could directly transfer English annotations onto the French track (leading to potentially noisier annotations) in order to train the French speaker identification module, we chose to automatically translate English transcripts into French before aligning them with French subtitles as described in Section 3.1.

4.2 Manual transcripts translation

To translate transcripts, we use NCODE, an open source n-gram Statistical Machine Translation (SMT) system[1]. This system achieved state-of-the-art performance in recent *Workshop on Statistical Machine Translation* (WMT) evaluation campaigns [9, 5]. NCODE implements the bilingual n-gram approach to SMT [11, 23, 14] that is closely related to the standard phrase-based approach [22].

In this approach, to translate a source sentence $\mathbf{s}$ into a target sentence $\mathbf{t}$, the translation process is decomposed into two steps: first the source sentence is reordered according to a set of rewriting rules so as to reproduce the target word order; this generates a word lattice containing the most promising source permutations. Then, this candidate lattice is translated in a monotonic way from left to right. The monotonic translation step involves to consider all the possible segmentation of the candidates in translation units (segments of contiguous words) and to propose several possible translations for these source segments. The best translation is then selected by maximizing the following inference term:

$$\underset{\mathbf{t},\mathbf{a}}{\operatorname{argmax}}\ p(\mathbf{t},\mathbf{a}|\mathbf{s}) = \underset{\mathbf{t},\mathbf{a}}{\operatorname{argmax}}\ \frac{1}{\mathcal{Z}_{\mathbf{s}}} \exp\Big(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{s},\mathbf{t},\mathbf{a})\Big) \quad (7)$$

[1] `http://ncode.limsi.fr`

where K feature functions (f_k) are weighted by a set of coefficients (λ_k), $\mathcal{Z}_{\mathbf{s}}$ is a normalizing factor, and $\mathbf{a}$ denotes the set of hidden variables corresponding to the reordering and segmentation of the source sentence. Since the translation step is monotonic, the peculiarity of this approach relies on the use of a n-gram translation model that estimates the probability of a sequence of bilingual units. Along with the n-gram translation model and a target n-gram language model, 13 conventional features are combined in Equation 7: 4 *lexicon models* similar to the ones used in standard phrase-based systems; 6 *lexicalized reordering models* [37, 15] aimed at predicting the orientation of the next translation unit; a "weak" distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system preference for short translations.

In the following experiments, we use the state-of-the-art system submitted to the WMT 13 campaign [5]. This large-scale system is fully described in [1] and was built using all the available data provided by the workshop organizers.

The French translated transcripts are then aligned with French subtitles as described for English in Section 3.1. Empirical results showed that the automatic translation were of reasonable quality. It is worth noticing that perfect translations are not required, but they must exhibit a sufficient word recall to obtain accurate alignments with subtitles.

5. EXPERIMENTAL SETUP

5.1 Evaluation corpus

Audio tracks, subtitles and manual transcripts are obtained from the publicly available `TVD` corpus [33] that provides all the necessary tools to generate these resources (either from physical DVDs of the series or from the Internet).

For evaluation purposes, reference annotations of the English audio tracks are obtained from previous work on the very same TV series [36, 4]. These reference annotations contains the speech turns of the five main characters of the series (*Sheldon*, *Leonard*, *Penny*, *Howard* and *Raj*) with all secondary characters grouped into a sixth class (*other*). Though they are not used here, it is worth mentioning that the *non-speech* regions are also segmented into several subclasses such as *music*, *silence* or *laughter*. Note that reference annotations are only available for the English audio tracks. Therefore, although the French track is used to improve the English one, the speaker identification approach cannot be directly evaluated for French alone.

5.2 Evaluation protocol

Experiments are conducted on the first six episodes of the first season of *The Big Bang Theory* TV series because manual speech turns annotations are only available for these very episodes. This amounts to a total duration of approximately two hours (each episode being twenty minutes long).

Due to the relatively limited size of the evaluation corpus, we opted for the leave-one-out cross-validation paradigm. Putting one episode aside, all other episodes are used to train both speech activity detection and speaker identification models. These models are then applied on the previously unseen test episode. This process is repeated for each episode and reported values are averaged over each run.

5.3 Evaluation metrics

Speech activity detection results are reported using three complementary evaluation metrics. We define the *Detection error rate* (DER) as the ratio of the duration incorrectly classified as *speech* or *non-speech* over the total duration of the episode. *Precision* is the ratio of the total duration reported as *speech* that is indeed annotated as *speech* in the reference annotation. *Recall* is the ratio of the total duration of *speech* according to the reference annotation that is is indeed detected as *speech* in the hypothesis.

Speaker identification results are reported using *Identification error rate* (IER), defined as follows:

$$\text{IER} = \frac{\text{miss} + \text{fa} + \text{confusion}}{\text{speech}} \tag{8}$$

where speech is the total duration of *speech* according to the reference annotation, miss (respectively fa) is the total duration of segments incorrectly classified as *non-speech* (resp. *speech*) and confusion is the total duration of *speech* segments whose detected label is incorrect. In other words, it is a compound metric that accounts for both speech turns detection and identification errors.

5.4 Implementation details

All three modules (speech activity detection, speech turn segmentation and speech turn identification) rely on Mel-Frequency Cepstral Coefficients (MFCC) features extracted every 16ms from a 32ms Hamming sliding window, using *Yaafe* open-source toolkit [25]. For speech activity detection, we use 12 MFCCs and energy first derivative, and 16 Gaussians with diagonal covariance matrix for each state (*speech* and *non-speech*). For speech turn segmentation, we use 12 MFCCs and energy, 1s-long left/right windows with a sliding step of 100ms. Speech turn identification relies on 13 MFCCs, their first and second derivatives and the energy first and second derivatives. The UBM is made of 256 Gaussians with diagonal covariance. Both GMM and HMM implementations are based on the *scikit-learn* toolkit [29]. Finally, bilingual fusion parameter α is set to 0.5 in our experiments – thus giving the same weight to both English and French tracks.

5.5 Reproducible research

Alongside the reproducible corpus, the source code necessary to reproduce and evaluate the results of all speaker identification experiments (including feature extraction, speech activity detection, speech turn segmentation and classification) is available as open-source software from the corpus webpage (`tvd.niderb.fr`).

6. RESULTS

6.1 Speech activity detection

The first set of experiments aims at showing that one can rely solely on DVD subtitles to train a speech activity detection module in a weakly supervised fashion. Table 1 compares the performance obtained by the fully supervised (*i.e.* trained using reference annotations) and weakly supervised (*i.e.* trained using readily available subtitles time spans) speech activity detection.

It shows that the latter achieves performance nearly as good as the former (8.1% *vs.* 7.8%), though they differ in

Approach	Subtitles	Supervised	Unsupervised
DER	19.8%	**7.8%**	**8.1%**
Precision	74.6%	94.8%	91.2%
Recall	95.6%	90.4%	94.1%

Table 1: Speech activity detection.

their behavior (better recall for the weakly supervised approach and better precision for the fully supervised one). Column *Subtitles* gives us a first explanation of why this is happening. It shows that subtitles cover most (95.6%) *speech* regions but also contain 25.4% of *non-speech* regions – therefore leading to a weakly supervised approach with an expected tendency to detect *non-speech* segments as *speech*.

6.2 Speaker identification

The second set of experiments focuses on speaker identification. Results are reported in Table 2 depending on whether identification is applied on *reference* speech turns (*i.e.* with perfect speech activity detection and speech turns segmentation) or speech turns obtained *automatically* via the fully/weakly supervised speech activity detection modules.

Speaker identification approach	Segmentation		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

Table 2: Speaker identification error rate (IER) with manual or automatic speech turn segmentation.

For a given speech turn, the *oracle* always projects onto it the correct reference label. If the reference label is a secondary character, then the oracle chooses the most frequent main character (here *Sheldon*). In case of ambiguity (*e.g.* when the speech turn covers more than one reference label), it is solved by choosing the reference label with maximum overlap duration. Therefore, its errors result either from secondary characters (for which no model is trained) or from segmentation errors (*i.e.* detected speech turns that actually cover speech turns of multiple characters). Its performance allows to estimate a lower bound of the impact of segmentation errors on other approaches. In particular, it shows that secondary characters account for only 10% of total speech duration and that incorrect speech activity detection and segmentation adds another 15.4% errors in total.

The *labeled subtitles* approach projects labeled subtitles (obtained in Section 3.1) onto the speech turns in the same way as the *oracle* does with reference labels. Its performance can be used as a measure of how noisy the data used for training the *weakly supervised* approach are. Moreover, its performance close to that of the *oracle* also indicates that, when subtitles and transcripts are available, one should use them directly instead of relying on automatic processing.

Finally, the *fully supervised* (resp. *weakly supervised*) approach rely on reference annotations (resp. automatically labeled subtitles) to train models for the five main characters. We conclude that it is not necessary to go through the costly process of manual annotation to train a speaker identification system. Indeed, relying on automatically obtained coarse annotations lead to the exact same overall error rates (19% for perfect segmentation and 38% for automatic segmentation). The p-value of 78% obtained in a *paired-samples t-test* confirms that the performance of the two approaches are not statistically different from each other.

6.3 Bilingual speaker identification

Finally, the last set of experiments is related to bilingual fusion as described in Section 4.1. Table 3 shows that bilingual (English and French) speaker identification significantly improves monolingual (English) speaker identification performance (down to 32.8% from 37.8%). A detailed analysis of the error rate shows that confusion errors on the five main characters are reduced from 7.1% to 2.1% – corresponding to a relative improvement of 70%. Most of the remaining errors are coming either from speech segmentation (15.4%) or secondary characters (15.3%) with no associated models.

	IER	Confusion	
		all characters	5 main char.
English	37.8%	22.5%	7.1%
Bilingual	32.8%	17.5%	2.1%
Improvement	−13%	−22%	**−70%**

Table 3: Bilingual speaker identification for $\alpha = 0.5$

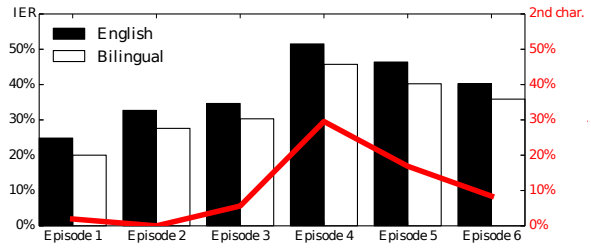

Figure 4: Performance breakdown per episode. Superimposed is the ratio of total speech duration uttered by secondary characters.

As highlighted by Figure 4, absolute improvement is consistent across all 6 episodes, on average 5% with a small standard deviation of 0.7%. Moreover, a *paired-samples t-test* (p-value $< 0.01\%$) confirms that the bilingual approach statistically and significantly outperforms the monolingual one. Additionally, note that the variation of performance between episodes is mostly explained by the ratio of total speech duration uttered by secondary characters for which no biometric models are available.

Though α was arbitrary set to 0.5 in the reported results, Figure 5 allows to better understand its influence on the overall performance. It shows that combining English and French approaches always outperforms the English-only system, whichever value is chosen for α. However, we notice the unexpected property that French speaker identification performs better than its English counterpart, even though the task is evaluated on the English track. This could mean that French actors' voices are easier to distinguish from each other than original actors' voices.

Figure 6 provides additional insight at the complementary information provided by the multilingual audio tracks.

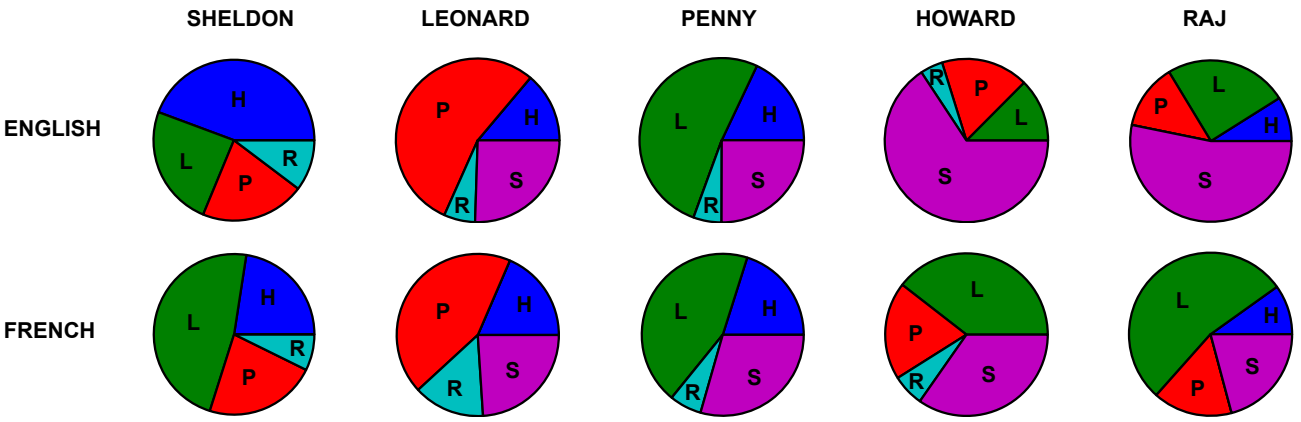

Figure 6: Distribution of confusion errors for each language and each main character (*Sheldon* (S), *Leonard* (L), *Penny* (P), *Howard* (H) and *Raj* (R)). For instance, the top-left pie chart shows that *Sheldon* is mostly mistaken for *Howard* by the English speaker identification module.

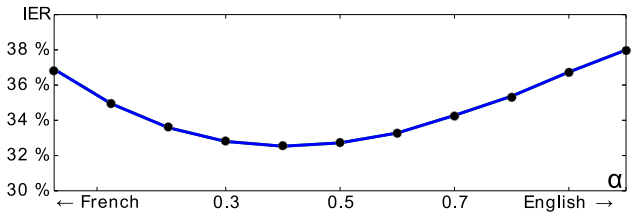

Figure 5: Influence of parameter α on the overall multilingual speaker identification performance.

While *Penny* – the only female character – is mostly confused with *Leonard* by both English and French systems, the English confusion patterns for *Sheldon*, *Howard* and *Raj* strongly differ from the French ones. In English *Sheldon* and *Howard* are mainly mistaken for each other, while in French both are mostly confused with *Leonard*. This confirms our initial motivation for multilingual speaker identification and explains why a simple fusion approach is able to circumvent monolingual confusions by leveraging complementary multilingual behaviors: two characters with similar voices in one language are unlikely to be confusable in the other language.

7. CONCLUSION

In this paper, we described a weakly supervised speaker identification system focusing on main characters of TV series. It relies on the combination of DVD subtitles and fan-made transcripts available on the Internet. It reaches the same overall performance as would a system trained using costly manual annotations (both for speech activity detection and speaker identification). Leveraging bilingual (English and French) audio tracks, we were able to further reduce main characters speaker identification errors from 7% to 2% (a 70% relative improvement).

Once the proposed system is trained, DVD subtitles and fan-made transcripts are no longer necessary to automatically detect and tag main characters speech turns as soon as a new episode airs. However, in case subtitles and/or transcripts are available, experimental results reported in Table 1 show that one should preferably rely on automatically *labeled subtitles* using the proposed alignment technique.

The two remaining main sources of errors of the proposed approaches are imperfect speech turn segmentation and missing speaker models for secondary characters (accounting for a total of 30% error rate). The proposed approach can be directly extended to secondary characters as long as they appear in at least one episode for which both subtitles and transcripts are available. However, we have yet to evaluate the influence of adding extra speaker models on the overall performance of speech turn identification. As far as speech turn segmentation is concerned, the shortness of speech turns and fast interactions between characters is a serious problem that needs to be addressed in the future.

Acknowledgments

This work was done in the context of the QCOMPERE project (funded by ANR) and the CHIST-ERA CAMOMILE project (funded by ANR, FNR and Tübitak).

8. REFERENCES

[1] A. Allauzen, N. Pécheux, Q. K. Do, M. Dinarelli, T. Lavergne, A. Max, H.-S. Le, and F. Yvon. LIMSI @ WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 62–69, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[2] C. Barras and J.-L. Gauvain. Feature and Score Normalization for Speaker Verification of Cellular Data. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 49–52, 2003.

[3] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multi-Stage Speaker Diarization of Broadcast News. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1512, 2006.

[4] M. Bäuml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised Learning with Constraints for Person Identification in Multimedia Data. In *International Conference on Computer Vision and Pattern Recognition*, 2013.

[5] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, 2013.

[6] H. Bredin. Segmentation of TV Shows into Scenes using Speaker Diarization and Speech Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, March 2012.

[7] H. Bredin, A. Laurent, A. Sarkar, V.-B. Le, S. Rosset, and C. Barras. Person Instance Graphs for Named Speaker Identification in TV Broadcast. In *Odyssey 2014, The Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014.

[8] H. Bredin and J. Poignant. Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, Lyon, France, August 2013.

[9] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, 2012.

[10] L. Canseco, L. Lamel, and J.-L. Gauvain. A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 415–419, 2005.

[11] F. Casacuberta and E. Vidal. Machine Translation with Inferred Stochastic Finite-State transducers. *Computational Linguistics*, 30(3):205–225, 2004.

[12] S. S. Chen and P. Gopalakrishnan. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.

[13] T. Cour, B. Sapp, A. Nagle, and B. Taskar. Talking Pictures: Temporal Grouping and Dialog-Supervised Person Recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[14] J. M. Crego and J. B. Mariño. Improving Statistical MT by Coupling Reordering and Decoding. *Machine Translation*, 20(3):199–215, 2006.

[15] J. M. Crego, F. Yvon, and J. B. Mariño. N-code: an open-source bilingual N-gram SMT toolkit. *Prague Bulletin of Mathematical Linguistics*, 96:49–58, 2011.

[16] Y. Estève, S. Meignier, P. Deléglise, and J. Mauclair. Extracting true speaker identities from transcriptions. In *Proceedings of the International Speech Communication Association*, pages 2601–2604, 2007.

[17] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" Automatic Naming of Characters in TV Video. In *British Machine Vision Conference*, 2006.

[18] G. Friedland, L. R. Gottlieb, and A. Janin. Joke-o-mat: browsing sitcoms punchline by punchline. *ACM Multimedia*, pages 1115–1116, 2009.

[19] J.-L. Gauvain and C.-H. Lee. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.

[20] V. Jousse, S. Petitrenaud, S. Meignier, Y. Estève, and C. Jacquin. Automatic Named Identification of Speakers using Diarization and ASR Systems. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Taïpei, Taïwan, April 2009.

[21] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly. A Presentation of the REPERE Challenge. In *International Workshop on Content-Based Multimedia Indexing*, pages 1–6, 2012.

[22] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[23] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R. Costa-Jussà. N-gram-based Machine Translation. *Computational Linguistics*, 32(4):527–549, 2006.

[24] A. F. Martin and M. A. Przybocki. The NIST 1999 Speaker Recognition Evaluation - An Overview. *Digital Signal Processing*, 10(1–3):1–18, 2000.

[25] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G.Richard. YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software. In *Proceedings of the 11th ISMIR Conference*, Utrecht, Netherlands, 2010.

[26] J. Mauclair, S. Meignier, and Y. Estève. Speaker Diarization : about whom the Speaker is Talking? In *IEEE Odyssey*, 2006.

[27] E. Myers. An O(ND) Difference Algorithm and its Variations. *Algorithmica*, 1(2):251–266, 1986.

[28] R. Nelken and S. Shieber. Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, 2006.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] J. Poignant, L. Besacier, V.-B. Le, S. Rosset, and G. Quénot. Unsupervised Naming of Speakers in Broadcast TV: using Written Names, Pronounced Names or Both? In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, Lyon, France, August 2013.

[31] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[32] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[33] A. Roy, C. Guinaudeau, H. Bredin, and C. Barras. TVD: a Reproducible and Multiply Aligned TV Series Dataset. In *LREC 2014, 9th Language Resources and Evaluation Conference*, 2014.

[34] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" - Learning Person Specific Classifiers from Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[35] J. Sivic and A. Zisserman. Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, 2009.

[36] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

[37] C. Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 101–104, 2004.

[38] S. E. Tranter. Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1013–1016, 2006.

[39] S. E. Tranter and D. A. Reynolds. An Overview of Automatic Speaker Diarization Systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, September.