TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Speech Activity and Speaker Change Point Detection for Online Streams

Summary of Dissertation

Study programme: P2612 – Electrical Engineering and Informatics
Study branch: 2612V045 – Technical Cybernetics

Author: **Ing. Lukáš Matějů**
Supervisor: Ing. Petr Červa, Ph.D.

Liberec 2019

Abstract

Speech Activity and Speaker Change Point Detection for Online Streams

The main focus of the thesis lies on two closely interrelated tasks, speech activity detection and speaker change point detection, and their applications in online processing. These tasks commonly play a crucial role of speech preprocessors utilized in speech-processing applications, such as automatic speech recognition or speaker diarization. While their use in offline systems is extensively covered in literature, the number of published works focusing on online use is limited. This is unfortunate, as many speech-processing applications (e.g., monitoring systems) are required to be run in real time.

The thesis begins with a three-chapter opening part, where the first introductory chapter explains the basic concepts and outlines the practical use of both tasks. It is followed by a chapter, which reviews the current state of the art and lists the existing toolkits. That part is concluded by a chapter explaining the motivation behind the thesis and the practical use in monitoring systems; ultimately, this chapter sets the main goals of the thesis.

In the thesis, the next two chapters cover the theoretical background of both tasks. They present selected approaches relevant to the work (used for result comparisons) or focused on online use.

The following chapter proposes the final speech activity detection approach for online use. Within this chapter, a detailed description of the development of this approach is available as well as its thorough experimental evaluation. This approach yields state-of-the-art results under low- and medium-noise conditions on the standardized QUT-NOISE-TIMIT corpus. It is also integrated into a monitoring system, where it supplements a speech recognition system.

The final speaker change point detection approach is proposed in the following chapter. It was designed in a series of consecutive experiments, which are extensively detailed in this chapter. An experimental evaluation of this approach on the COST278 database shows the performance of approaching the offline reference system while operating in online mode with low latency.

Finally, the last chapter summarizes all the results of the thesis.

Keywords: Deep Neural Networks, Online Streams, Speech Activity Detection, Speaker Change Point Detection, Weighted Finite-State Transducers.

Contents

Introduction

Nowadays, an increasingly overwhelming amount of audio data is produced every day by various media streams (television, radio, etc.) as well as many other sources (e.g., the Internet). Unfortunately, most of this data lacks labels (annotations, tags) of any kind that would be useful for a wide range of applications; in this case, for speech processing. The aforementioned labels vary greatly; they can, e.g., include speech transcription, subtitles, change of speaker, or name of the played song, to name a few. They can even carry time stamps, which can be further utilized for audio searching, indexing, or data retrieval. Speech Activity Detection (SAD) and Speaker Change Point (SCP) detection (often called speaker segmentation) are among the tasks that can create such labels. The former is a task of identifying and labeling speech and non-speech segments in an utterance while the latter, for a given utterance, finds and labels changes between different speakers (i.e., it is a task of detecting exact moments when a change of speaker occurs). As their output, both of these tasks split the recording into segments (speech/non-speech or speaker-homogeneous) and provide start- and end- time stamps of these newly defined blocks.

In general, SAD and SCP detection are closely interrelated tasks. As such, they form an integral preprocessing component of many speech processing applications including, e.g., speaker verification and identification, language, gender or emotion detection, audio indexing and retrieval, or automatic speech transcription. Specifically, in speech transcription, implementation of SAD can significantly speed up the processing as well as increase the overall performance as the non-speech segments are omitted from transcription. Finally, SAD usually plays the role of the preprocessor even for SCP detection, which is only run on obtained speech segments.

SCP detection, in conjunction with speaker clustering, results in a speaker diarization system. Speaker diarization focuses on answering the question "who spoke when?" (it breaks down the recording into speaker-homogeneous segments and clusters the segments according to the speaker's identity), and it can be further extended into speaker verification and identification systems. The research is driven by challenges held by the National Institute of Standards and Technology (NIST). Additionally, SCP detection can be employed for tasks such as rich transcription, dialog detection, speaker tracking, multi-speaker detection, and more. Lastly, the extracted speaker-homogeneous segments can also be used as training data for speaker-adaptive approaches to Automatic Speech Recognition (ASR).

The diverse applications of SAD and SCP detection make both of these tasks popular research topics. Numerous research groups and research centers compete worldwide and propose novel approaches in pursuit of improving the state-of-the-art results. Challenges are also being held quite regularly. The popularity of these research topics can also be documented by large amounts of papers accepted at international conferences, such as Interspeech or ICASSP. With the recent boom in deep learning in mind, SAD and SCP detection attract more and more researchers every day, and much exciting work is being published every year.

1 State of the Art

At present, SAD and SCP detection are generally treated as machine learning tasks. Recently, deep learning has extensively been applied to both of these tasks to improve their performance, and subsequently the results achieved. Both of these tasks are usually performed in two consecutive phases: feature extraction and classification. Moreover, both can be run in an offline or online mode. In the former mode, no additional restrictions are applied, and low latency and real-time processing are not vital. However, they become crucial in the latter mode. Furthermore, an online decoder may only perform one left-to-right pass through the input data. These additional restrictions result in a limited amount of published work for online use.

1.1 Speech Activity Detection

As already stated above, the majority of the existing SAD approaches operate in two subsequent phases: feature extraction and speech/non-speech classification. In the former phase, the classic approaches for feature extraction utilize energy [1], zero-crossing rate [2] or auto-correlation function [3]. The family of more complex features, which have also been successfully applied, includes Mel-Frequency Cepstral Coefficients (MFCCs) [4, 5], pitch related features [6], multi-band long-term signal variability features [7] or i-vectors [8]. Bottleneck (BTN) features extracted from Deep Neural Networks (DNNs) have also been proposed [9, 10].

In the latter phase, various classification algorithms can be used, such as support vector machines [11] or Gaussian Mixture Models (GMMs) [12–14]. In recent years, various DNN architectures have been frequently employed, including fully connected feed-forward DNNs [4, 15, 16], Convolutional Neural Networks (CNNs) [17, 18], dilated CNNs [19] or Recurrent Neural Networks (RNNs) [20–22]. More complex approaches, such as jointly trained DNNs [23], boosted DNNs [24] or a combination of DNNs and CNNs [25], have also been proposed. Furthermore, an adaptive context attention model was suggested in [26]. The output from a given classifier can also be smoothed to further improve the accuracy of the detection. Over the years, various techniques, such as the Viterbi decoder [4] or Weighted Finite-State Transducers (WFSTs) [27], have been applied for this purpose.

Most of the previously mentioned works primarily aim at offline application, or the focus is not specified in the given publications. The limited amount of approaches developed namely for the online task include, for example, conditional random fields [28] or accurate endpointing with expected pause duration [29]. An unsupervised approach to real-time Voice Activity Detection (VAD) was introduced in [30, 31]. Another approach in [32] utilizes short-term features. Recently, a causal voice activity detector based on DNNs has been suggested in [33]. In [34], an online speech activity detector using simultaneously trained neural networks is shown. Finally, the authors of [35] studied the impact of lowering the representation precision of DNN weights and neurons on the accuracy and delay of VAD.

1.2 Speaker Change Point Detection

In the literature, SCP detection commonly utilizes SAD as a preprocessor, and it is thus carried out only on speech segments. Furthermore, it is usually done without any prior knowledge about the identity or even the number of speakers in the recording (i.e., it is treated as a speaker-independent task). Similar to SAD, most of the existing SCP detection approaches are designed in two consecutive phases: feature extraction and change point detection itself.

In the first phase, various types of input features have been applied over the years. In the early years, more straightforward ones were successfully employed, such as zero-crossing rate or pitch [36]. MFCCs [37, 38] were probably the most commonly used features, followed by line spectrum pairs [39]. Recently, the main focus has shifted to crafting more complex features capturing more speaker-specific information. Nowadays, i-vectors [40, 41] are the go-to features for most state-of-the-art systems. Alternatively, DNNs have also been successfully utilized to extract complex features [42, 43]. Furthermore, d-vectors were presented in [44], yielding excellent results. The latest trend goes in the direction of deep speaker embeddings [45–48] designed for end-to-end systems.

In the second phase, the SCP detection approaches can be divided into three main categories: metric-, model- and hybrid-based. The first type requires a distance metric to be defined first. After that, usually, two adjacent windows are shifted alongside the recording, and the distance between them is computed. If the distance is larger than a predefined threshold, a change point is detected. The most commonly used distance metrics include the Bayesian Information Criterion (BIC) [49–51], the generalized likelihood ratio [52], the Gaussian divergence [53], the Kullback-Leibler divergence [54], or one-class support vector machines [55]. A model-based approach utilizes trained models from labeled audio data to detect speaker change points. Among the most common approaches, there are the Hidden Markov Models (HMMs) [56], the GMMs [57], and the eigenvoice-based models [58]. Deep learning approaches based on DNNs [43, 59], CNNs [60, 61], unidirectional [62], or bidirectional [63, 64] long short-term memory RNNs all yield excellent results.

Most of the approaches cited so far were designed with regard to the best possible quality of detection, and all of them are, of course, applicable to offline processing. However, the earlier discussed restrictions of online application are usually not taken into account during design, and the usability of these methods for online mode is therefore limited (or not discussed in the respective papers). That means that the number of approaches explicitly designed for real-time processing is much smaller. In the early years, an online SCP detector utilizing the Bayesian fusion method was proposed [65, 66]. Other works focused on BIC [67, 68], XBIC [69], log-likelihood ratio [70], or GMMs [57, 71–73]. In [74], the authors explored BIC, i-vectors, and within-class covariance normalization for speaker diarization. The use of i-vectors for diarization was also investigated in [75]. Features extracted from neural network were explored in [76]. Finally, the authors in [77] studied in detail the influence of the online environment of several SCP detection approaches on a diarization system.

2 Motivation and Goals

A detailed examination of the current state of the art in speech activity detection, as well as speaker change point detection, reveals two prominent features: a) deep learning is pushing the field further; and b) there is a significant lack of online SAD and SCP detectors. With this information, it is feasible to set up the motivation and consequently, the main goals of the thesis.

2.1 Motivation

Over the past few years, significant breakthroughs [78] have been achieved in deep learning. These breakthroughs have resulted in many novel approaches in various research fields, such as speech recognition [79–81], visual object recognition [82, 83], natural language processing [84, 85], and more, all yielding excellent results as compared with the previously used conventional techniques. These successes have understandably led to further application of deep neural networks to a much more varied range of research tasks. In this case, deep learning is applied to speech activity detection and speaker change point detection. Lately, several papers dealing with this topic have been published for both tasks, yet there is a lot of room for further experimentation, tuning up, and improvements. Performance in the online mode, especially, can be further enhanced.

Speech activity detection and speaker change point detection represent a very active research topic due to their varied use in a wide range of speech processing applications. Over the years, most of the published works have strictly focused on the offline use as it allows more freedom during the design of the detector. It is also easier to tune the performance of an offline system to achieve excellent results (i.e., multiple passes through data, processing of whole recording, a fusion of methods, etc.) than its online counterpart. Moreover, for many applications, it is a perfectly viable and even preferred solution. However, some applications (e.g., Television and Radio [TVR] monitoring systems) need to operate in real time and with low latency. These additional restrictions usually result in somewhat limited performance. Extension of the existing offline methods to their online use is a commonly cumbersome and complicated process, which is even quite often impossible. Moreover, the performance is usually affected as well. When designing an approach that may be used in a real-time application, it is generally more convenient to circumvent these restrictions from the initial stages of development. Online speech activity detection and speaker change point detection approaches (based on deep learning) that would reach results at least comparable with their offline counterparts would be very beneficial for many real-time speech processing applications (e.g., TVR monitoring system) in both commercial and research spheres (i.e., they could push the field further).

2.2 Practical Use in TVR Monitoring System

The author's lab has been focusing on speech processing and ASR for a long time. The TVR monitoring system developed at SpeechLab@TUL in cooperation with the NanoTrix company carries out 24/7 online transcription of radio and TV broadcasts in various languages. In the peak hours (during the day), it transcribes up to 120 streams in parallel in real time. During the non-prime hours (mostly at night), it still processes at least 20 online streams every second. The daily average ranges from 60 to 80 simultaneously transcribed online streams. Approximately "133" days (3,196 hours or 750 GB) of recordings are being processed every day.

Integration of SAD and SCP detection approaches into this existing system would be beneficial for many reasons. First, SAD would be used as a preprocessor for online streams to filter out non-speech events and run the transcriber only on speech ones. This should result in a significant reduction of processing time, and it should ease the CPU load as well (if the stream contains a lot of non-speech segments, e.g., music stream radios). It should also yield a better accuracy of transcriptions as the non-speech parts are omitted from being transcribed (i.e., less gibberish). Furthermore, the obtained speech segments would be used as inputs into the SCP detection and potentially other speech processing applications.

Second, the SCP detector would find and label transitions from one speaker to another. These newly defined labels would ease the handling of online streams as they would provide additional information about the content. They would also segment the streams into smaller speaker-homogeneous chunks, which could easily be further utilized. These chunks form a starting point for a full diarization system, which could be extended to speaker verification and identification systems to provide the transcribed streams with even more valuable information. The final detected segments could also be extracted and used as training data for future speaker-adaptive approaches to speech recognition.

2.3 Goals

The main goals of the thesis are thus to:

I. develop speech activity detection approach and speaker change point detection approach that:
    1. utilize state-of-the-art techniques, specifically including DNNs;
    2. allow for robust speech/non-speech and speaker change point detection;
    3. operate in an online mode with low latency in order to process real-time streams;
    4. can be integrated into the existing TVR monitoring system developed at the author's lab in cooperation with the NanoTrix company;

II. verify the proposed approaches and compare their results on publicly available datasets with selected existing approaches/toolkits.

3 Proposed Speech Activity Detection Approach

The final approach to speech activity detection was proposed in a series of consecutive experiments, all described and heavily discussed within this chapter. The majority of this designing process was covered in [86–88], and portions of the respective papers were directly utilized in the thesis. This chapter thus describes evaluation metrics, training and development data, experimental evaluation of all steps taken, evaluation on standardized QUT-NOISE-TIMIT [89] corpus, evaluation in real speech transcription system, and at last, it sets the final SAD approach.

3.1 Evaluation Metrics

In total, seven different commonly utilized metrics were employed for the evaluation of SAD. These metrics can be grouped into three main subsets: overall accuracy metrics, change point quality metrics, and performance metrics.

Overall Accuracy Metrics

The main focus of this group of metrics is the accuracy of speech and non-speech segments on a frame-level (i.e., the recording is treated as a sequence of speech and non-speech frames). In this case, each frame is considered independent, and only a direct comparison between the reference frame and the corresponding decoded frame (frame pair) is evaluated. If the frame pair is matched, it is considered as a hit; otherwise, it is a miss. For this task, four closely related metrics were applied.

The first metric, Frame Error Rate (FER), is defined as a ratio of non-matching frame pairs to all frames in reference. Miss Rate (MR), the second metric, explores only the speech segments. It can be expressed as a ratio of speech frames misclassified as non-speech ones to all speech frames in reference. False Alarm Rate (FAR) is defined analogously to MR but for non-speech frames [4]. Finally, Half-Total Error Rate (HTER) can be defined as an equal-weighted average of MR and FAR.

The optimal SAD approach should minimize the miss rate while keeping the false alarm rate relatively low. The reason is that the following speech processing system (e.g., SCP detector or speech transcriber) should get all speech frames possible with only a limited amount of non-speech events added.

Change Point Quality Metrics

Change point quality metrics offer an alternative view on the performance of SAD. Instead of a frame-based evaluation, they explore the recording as a sequence of consecutive speech and non-speech events, and more specifically, as the name suggests, they focus on the accuracy of detected (computed) change points between these events. For this task, two distinct metrics, F-measure and $\delta_{2/3}$, were employed.

To define these two metrics, the detected and the reference change points have to be aligned at first, e.g., by the bidirectional search for the nearest neighbor [90]. After the alignment, the matched detected and reference change points are labeled as hits, while the errors are marked as insertions (when detected change point does not match any of the reference change points) and deletions (when reference change point is not matched by any of the detected change points).

Given the values of hits, insertions and deletions, Precision (P) and Recall (R) can be expressed. Precision is defined as a ratio between the number of correctly detected change points and the number of detected change points, while recall is expressed as a ratio between the number of correctly detected change points and the number of change points in reference. Precision and recall are in a contradictory relationship with each other (i.e., when one improves the other one worsens). For this reason and to express the performance with only one value, F-measure is defined:

$$F - measure[\%] = \frac{2 * R * P}{R + P} \quad . \tag{3.1}$$

Given the correctly detected change points (hits), it is also possible to calculate an error value for each hit (in seconds) and sort all the hits according to these values in ascending order. In this work, $\delta_{2/3}$ was utilized. It expresses (in seconds) the maximal error of the alignment for the first two-thirds of the sorted (best) hits.

Performance Metrics

The last set of metrics monitor the performance of SAD in an online environment. Two different metrics, Latency (L) and Real-Time Factor (RTF), were utilized. The former one is defined as an average time between the detected change point, and the moment the decoder outputs the change point label. The latter metric expresses the speed of decoding as a ratio of processing time to the duration of the recording.

3.2 Data Used

For training, in total, 67 hours of recordings have been gathered and utilized. The speech is represented by 30 hours of clean speech recordings of English and several Slavic languages (Czech, Slovak, Polish, Russian, and Croatian). These recordings originally served as training data for speech transcription systems. The non-speech is modeled by 30 hours of music of different genres with the addition of 7 hours of non-speech events/noises. Lastly, the annotations were done automatically, speech label for clean speech utterances and non-speech one for everything else.

The data used for development consists of 6 hours of TV and radio recordings in several Slavic languages (Czech, Slovak, Polish, and Russian). It contains not only clean speech segments but also segments with music, background noises, jingles, and advertisements. Annotations of this data were obtained in a two-step process. At first, speech/non-speech labels were produced automatically by the baseline DNN-based approach introduced in Sect. 3.3. These obtained labels were then corrected and fine-tuned by hand. In total, 70% of all frames are marked as speech ones.

3.3 Baseline DNN-Based Approach

The baseline speech activity detection approach employed a feed-forward deep neural network with a binary output (speech or non-speech) as a classifier (i.e., without any smoothing). The DNN had 5 hidden layers, each consisting of 128 neurons. The ReLU activation function and mini-batches of size 1024 were used within 10 epochs of training. The learning rate was set to 0.08. 39-dimensional log Filter Bank Coefficients (FBCs) were used as features. The input vector for DNN had a length of 51 and was formed by concatenating 25 previous frames, the current frame, and 25 following frames. Local normalization was performed within one-second windows.

The performance of the baseline approach is summarized in Table 3.1 (see its first row). It is evident that it missed approximately 4% of speech segments. This fact affects the accuracy of the possible speech transcription system negatively, as the segments incorrectly marked as non-speech would not be transcribed. Another problem of the baseline detector was the time precision of the change-point detection: the achieved value of $\delta_{2/3}$ was 0.42 seconds. This is also due to the fact that it is sometimes hard even for human annotators to determine the exact frame where a state change occurs. The baseline detector also produced a high number of false non-speech segments with a very short duration of one or two frames.

Table 3.1: Summarized results of the proposed SAD approach.

approach	FER [%]	MR [%]	FAR [%]	F [%]	$\delta_{2/3}$ [s]
baseline DNN-based	4.7	3.7	7.1	0.3	0.42
+ basic smoothing	2.9	2.2	**4.7**	28.5	0.27
+ artificial training data	3.1	**0.3**	10.1	41.3	0.34
modified artificial data + context-based smoothing	**2.4**	0.5	7.2	**52.7**	**0.26**

Note that the presented DNNs for all SAD experiments were trained on GPU using the torch framework[1]. The training scripts are available at the author's GitHub[2].

3.4 Smoothing the Output from DNN

As mentioned in the previous section, the baseline detector classified every input frame independently. On the other hand, every speech or non-speech segment usually lasts for at least several frames. Therefore, the following efforts were focused on smoothing the output from the DNN. For this purpose, weighted finite-state transducers were utilized using the OpenFst library[3].

[1] http://torch.ch/
[2] https://github.com/1shark1/nnet/
[3] http://www.openfst.org/twiki/bin/view/FST/WebHome

The resulting scheme consists of two transducers. The first models the input signal (see Figure 3.1). The other one is the transduction model and represents the smoothing algorithm (see Figure 3.2). It consists of three states. The first state, denoted by 0, is the initial state. The transitions between states 1 and 2 emit the speech/non-speech labels and are penalized by penalty factors P1 and P2, respectively. Their values (500 and 500) were tuned on a different dataset.

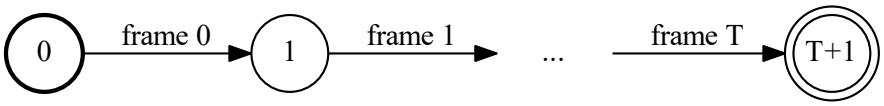

Figure 3.1: A transducer modeling the input signal for SAD.

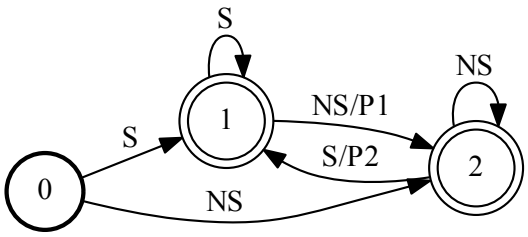

Figure 3.2: A transducer representing the basic smoothing model for SAD.

Given the two described transducers, the decoding process is performed using the on-the-fly composition of the transduction and the input model of unknown size. This is possible since the input is considered to be a linear-topology, unweighted, epsilon-free acceptor. After each composition step, the shortest-path (considering tropical semi-ring) determined in the resulting model is compared with all other alternative hypotheses. When a common path is found among these hypotheses (i.e., with the same output label), the corresponding concatenated output labels are marked as the final fixed output. Since the rest of the best path is not known with certainty, it is denoted as a temporary output (i.e., it can be further refined).

The results obtained with the aid of the DNN-based approach with smoothing are summarized in the second row of Table 3.1. They show an overall significant boost in performance. For example, F-measure improved from 0.3% to 28.5%, MR was reduced from 3.7% to 2.2%, and the value of $\delta_{2/3}$ improved noticeably from 0.42 seconds to 0.27 seconds.

3.5 Using Artificial Training Data

The level of MR yielded so far, i.e., around 2%, would still lead to a small increase in the Word Error Rate (WER) of a transcription system (e.g., from 13% to 14%), as the speech frames incorrectly classified as non-speech would be omitted from transcription. Upon closer inspection, the detector specifically misclassified the speech segments with background noise. The reason for this behavior is that the

speech data used for DNN training so far were recorded only in clean conditions (i.e., without any background noise).

Hence in the next step, the goal was to employ training data containing non-speech events, such as music or jingles in the background. Due to the lack of such annotated data, an artificial dataset created by mixing 30 hours of clean speech with non-speech recordings was constructed. For this purpose, a larger set of non-speech recordings of a total length of 100 hours was prepared first. After that, every speech recording was mixed with a randomly selected non-speech recording from the prepared set. Note that every non-speech recording used for mixing had to have the same or longer duration than the given input speech recording (the selected non-speech recording was trimmed to match the length of the speech recording) and its volume was increased or decreased to match the desired level of signal-to-noise ratio (which was also selected randomly from an interval between −30 dB and 50 dB).

The labeling of this artificial data was carried out automatically: when the SNR of the recording was higher than a defined threshold of 0 dB, the recording was marked as speech. In the opposite case, the recording was labeled as non-speech.

The results after using only these 30 hours of mixed training data are shown in the third row of Table 3.1. It is evident that this approach led to an increase in F-measure and a significant reduction in MR from 2.2% to 0.3%. Unfortunately, these improvements are all accompanied by an increase in FAR and, even more importantly, an increase in $\delta_{2/3}$ from 0.27 seconds to 0.34 seconds. Due to these issues, a further refinement of the smoothing algorithm was investigated.

3.6 Improved Context-Based Smoothing

The proposed refinement of the smoothing scheme is depicted in Fig. 3.3. In this case, both the speech and non-speech events are represented as sequences of three states, where the first and third states (the outer states) model the context. Similarly to smoothing without any context, the penalties are defined just for transitions between the speech and non-speech events, i.e., for transition a) from the end state of speech (*stop_S*) to the start state of non-speech (*start_NS*), and b) from the end state of non-speech (*stop_NS*) to the start state of speech (*start_S*). Their values were fine-tuned on a different dataset.

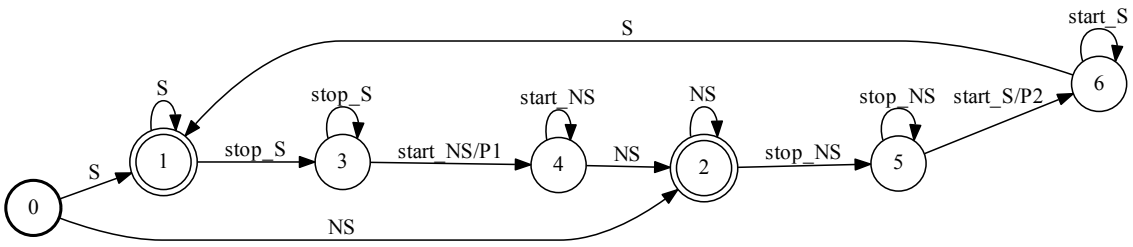

Figure 3.3: A transducer representing the context-based smoothing model for SAD.

To prepare training data containing transitions between speech and non-speech events, the dataset from Sect. 3.5 was modified. At first, two recordings were chosen

randomly from the artificial training set: one speech and one non-speech. After that, these two recordings were joined in random order. The resulting recording then contained one of the two possible transitions (i.e., from speech to non-speech or from non-speech to speech) and was annotated automatically as follows:

1. The number of transition frames was derived from the input feature context window (25-1-25).

2. Only the 50 frames at the inner boundary of the two joined recordings were annotated as transitional, i.e., using 25 labels *stop_S* followed by 25 labels *start_NS* or 25 labels *stop_NS* followed by 25 labels *start_S*.

3. All other frames were marked as either speech or non-speech.

Finally, the last change associated with the integration of context-based smoothing lies in the DNN model. Instead of the original two output neurons, there are now 6 (speech, non-speech and 4 transitional ones: *start_S*, *stop_S*, *start_NS*, *stop_NS*) to match the smoothing scheme and annotation style of data.

The results of the experiment with the context-based smoothing (see the fourth row of Table 3.1) show that this approach addresses the issue of an increase in $\delta_{2/3}$, which has emerged due to the use of the artificial training data (see the third row of Table 3.1). The value of $\delta_{2/3}$ was reduced from 0.34 seconds to 0.27 seconds. At the same time, a significant decrease in the FAR, an increase in F-measure, and only a slight decrease in MR by 0.2% were achieved when compared to the previous experiment. After scoring these results, the proposed approach was considered final.

3.7 Online Performance

An online performance of the proposed SAD approach was closely monitored throughout the whole design and experimental evaluation. This performance is crucial for the approach to be integrated into the target TVR monitoring system. The proposed approach averaged RTF of 0.01 and 2-second latency. Note that Intel Core i7-3770K @ 3.50GHz was used for the computations. The achieved performance is well suited for seamless use in real-time processing without any major delay.

3.8 Evaluation on QUT-NOISE-TIMIT Corpus

So far, all of the experiments were conducted only using the development dataset, which was designed explicitly within the thesis. That is not suitable for comparison purposes because the dataset has not been used anywhere else or even released to the general public. To compare the proposed approach with different SAD approaches presented in the literature, the QUT-NOISE-TIMIT [89] corpus was employed.

The QUT-NOISE-TIMIT corpus was designed for training and testing of various SAD approaches under different SNR conditions. For this purpose, the authors gathered background noises across 5 unique scenarios (cafe, car, home, reverb, and

street) and mixed them with a clean speech from TIMIT corpus [91] creating new recordings (i.e., the QUT-NOISE-TIMIT corpus) with varying amount of speech, length (60 or 120 seconds) and SNR level (−10, −5, 0, 5, 10 or 15 dB).

The authors also provided an evaluation protocol. It states that during training, the only prior knowledge given to the system is the SNR level of the target environment: low noise (10, 15 dB), medium noise (0, 5 dB), or high noise (−10, −5 dB). After the decoding is done, the final speech/non-speech segments are aligned with QUT-NOISE-TIMIT ground truth labels, and MR, FAR, and HTER are evaluated.

The evaluation on the QUT-NOISE-TIMIT corpus shows the performance of the proposed approach in comparison with five approaches already presented in [89] and two techniques reaching the state-of-the-art results [14, 92]. The five approaches are: standardized VAD system ITU-T G.729 Annex B [93], standardized advanced front-end ETSI [94, 95], Sohn's likelihood ratio test [96], Ramirez's long-term spectral divergence [97] and GMM-based approach with the use of MFCCs [89]. The latter two techniques are voice activity detection using subband noncircularity [92] and complete-linkage clustering for VAD [14].

Figure 3.4 presents the results of this comparison under low-, medium- and high-noise conditions. As the results show, the proposed approach outperformed all other systems by a fair margin under low- and medium-noise conditions. The absolute reduction in the HTER was more than 2% over the formerly best complete-linkage clustering. The exact achieved values of the HTER were 2.6% and 5.8% under low- and medium-noise conditions, respectively. Under high-noise conditions, the complete linkage clustering approach surpassed all other systems, including the proposed SAD approach (by approximately 2%). However, the proposed approach still outperformed all other systems (by at least 10%). In conclusion, the proposed SAD approach yielded state-of-the-art results under low- and medium-noise conditions.

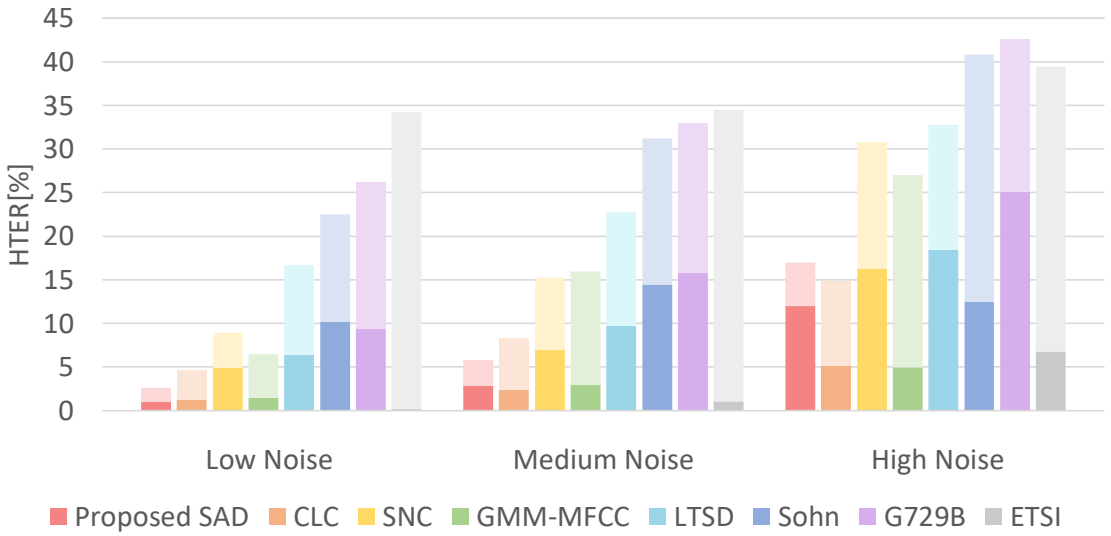

Figure 3.4: A comparison of the proposed approach with other systems under low-, medium- and high-noise conditions (QUT-NOISE-TIMIT corpus). The contribution of MR and FAR to HTER bars is displayed by darker and lighter shades, respectively.

3.9 Evaluation in Real Speech Transcription System

Given the findings and results from all previous experiments, the final proposed SAD approach was integrated into the TVR monitoring system developed at the author's lab in cooperation with the NanoTrix company and thus evaluated in a real speech transcription system.

Four metrics were applied to evaluate the performance of speech transcription. The first three, WER, Word Accuracy (WAcc) and Percent Correct (PC), focus on the quality of transcriptions, while the RTF evaluates the real-time performance.

For evaluation, two datasets of Czech broadcasts have been utilized. The first dataset represents 4 hours (22,204 words) recorded from a Czech live news TV channel. Approximately 60% of its content consists of speech segments. The length of the other dataset is 8 hours, it contains 7,212 words, and speech frames form only 10% of its content. This dataset represents a broadcast of a Czech radio station.

The transcription system employed an acoustic model based on an HMM-DNN hybrid architecture [79], where the baseline GMM was trained as context-dependent, speaker-independent and contained 3,886 physical states. 270 hours of clean speech were employed for training. The DNN hyper-parameters were derived from [98]. The input features were 39-dimensional FBCs, and the input feature vector had a length of 11 frames. The linguistic part of the system was composed of a lexicon and a language model. The lexicon contained 550,000 entries with multiple pronunciation variants, and the language model was based on bigrams.

Within the performed experiments, both evaluation datasets were transcribed a) with and b) without the use of the proposed SAD approach. The results are presented in Table 3.2. They reveal that the utilization of the proposed approach was advantageous on both evaluation datasets. The yielded PC and WER (WAcc) show that SAD limited the insertions coming from the non-speech parts and omitted hardly any speech parts. The proposed approach allowed the transcription system to operate with improved accuracy and, at the same time, RTF was almost two times, and more than ten times lower for the first and second evaluation datasets, respectively. Of course, the reason for this difference is that the data in the second dataset contains fewer speech segments. Finally, the latency was around 2 seconds. In conclusion, the transcription system complemented with the proposed SAD approach can be utilized for online speech transcription without any major delay.

Table 3.2: An evaluation of the proposed approach in a speech transcription system.

dataset	SAD	WER [%]	WAcc [%]	PC [%]	RTF
live news TV channel	yes	**12.4**	**87.6**	**89.7**	**0.42**
	no	12.7	87.3	**89.7**	0.77
local radio station	yes	**14.0**	**86.0**	**88.5**	**0.08**
	no	17.9	82.1	88.4	0.83

4 Proposed Speaker Change Point Detection Approach

Inspired by the proposed SAD approach, the final SCP detection approach was proposed in several successive experiments heavily detailed within this chapter. This development was published in [99], and portions of the paper were reused in the thesis. Ultimately, this chapter describes the evaluation metrics, training, development and evaluation data, experimental evaluation of all steps taken, evaluation on the COST278 [100] database, and finally, it sets the final SCP detection approach.

4.1 Evaluation Metrics

The evaluation metrics for SCP detection were close to identical to the ones used for SAD due to the similarity of both tasks. The overall accuracy metrics are the only exception because framewise evaluation is not particularly valuable for change point detection (i.e., the main concern is the actual placement of speaker transitions). Therefore, the metrics for SCP detection can be divided into two subsets: change point quality metrics and performance metrics. In total, 6 metrics were observed.

For the former subset, four metrics, specifically precision, recall, F-measure, and $\delta_{2/3}$, were employed. Precision and recall were additionally reported to provide more information about the errors the decoder makes (i.e., falsely detected change points result in worsened precision while undetected change points yield worse recall). The latter group consists of two previously introduced metrics: latency and RTF.

4.2 Data Used

For training, 20,000 recordings, each with an average length of 5 seconds, have been prepared with the help of automatic Czech TV/radio broadcast transcriptions. Each of these recordings contains exactly one speaker change point (i.e., the set consists of 20,000 speaker transitions). These transitions can be divided into four distinct groups (female to female, female to male, male to female, and male to male). Each of them is represented by 5,000 change points. Note that each recording was extracted from a whole utterance, and there are no artificial cuts or changes in channels.

The annotations of this data were generated in a fully automated way. The frame corresponding to the actual change point, as well as the safety collar frames around it, were labeled as change points. This safety collar was set to 1 second (100 frames), i.e., 50 frames before and 50 frames after the actual change point were considered as speaker transition frames. That is due to the fact that a) determining the precise change point is quite often an ambiguous task (silence, crosstalk, etc.), and b) it is necessary to provide DNN training with enough information about the speaker transitions. The remaining frames were labeled as no change point.

For development purposes, the Czech train subset of standardized COST278 [100, 101] pan-European broadcast news database has been utilized. Accurate annotations are provided by the database. For evaluation, the Czech test subset of COST278 has been employed. It consists of four recordings of different Czech broadcasts (ČT1, Nova and Prima) in a total length of 90 minutes. It contains not only clean speech segments but also segments with background noise and jingles. In total, 379 speaker change points are labeled within the data.

4.3 Reference Results

To obtain reference results with an offline system, publicly available LIUM Speaker Diarization toolkit [102, 103] was used. The SCP detection portion of the system is covered by BIC segmentation and BIC clustering, followed by segmentation based on Viterbi decoding and boundary adjustments. The system is also supplemented with a pre-trained model fine-tuned for TV and radio broadcasts. During the evaluation, the LIUM toolkit was operated with an RTF of 0.016, achieving reference results in F-measure of 84.6% and $\delta_{2/3}$ of 0.13 seconds (see the first row in Table 4.1).

Table 4.1: Summarized results of the proposed SCP detection approach.

approach	P [%]	R [%]	F [%]	$\delta_{2/3}[s]$	RTF	L [s]
LIUM toolkit	**89.9**	80.0	84.6	**0.13**	**0.016**	-
DNN + WFST decoder	59.4	63.6	61.4	0.24	0.022	2.4
+ enhanced data	67.0	70.7	68.8	0.21	0.022	2.3
+ Δ MFCC	72.8	74.7	73.7	0.19	0.024	**1.9**
+ CNN	79.3	77.8	78.6	0.17	0.054	**1.9**
+ 2.5-second context window	80.3	81.8	81.1	0.17	0.054	2.3
+ 1-second long transition model	82.7	81.8	82.2	0.17	0.065	2.9
+ tuned for offline use	86.7	**84.4**	**85.6**	0.18	0.079	4.8

4.4 Initial Approach Based on DNN and WFST

The initial SCP detection approach was inspired by the proposed SAD approach designated for online use. This SCP detection approach was based on DNN trained as a binary classifier (change point or no change point) and WFST designed as an online decoder detecting speaker transitions given the output from the DNN.

The binary DNN was trained using the following hyper-parameters: 2 hidden layers with 64 neurons per layer, the ReLU activation function, a learning rate of 0.08, mini-batches of size 1024, and 15 epochs. 39-dimensional MFCCs were employed for the feature extraction. The input feature vector was formed by concatenating 100 previous frames, the current frame, and 100 following frames (i.e., a 2-second

context window). No local normalization was applied. Note that all the DNNs for all SCP detection experiments were trained on GPU using the PyTorch framework[1].

As stated above, WFSTs were utilized (using the OpenFst library) as an online decoder. The decoding scheme consists of two transducers. The first one models the input signal (see Fig. 4.1), while the second one is the transduction model and represents the change point detection (see Fig. 4.2). It consists of two states, 0 and 1. The transitions between states 0/1 emit labels the start/end change points. The resulting change point is placed in the middle between these two labels. The transitions are also penalized by factors P1 and P2, whose values were fine-tuned on the development set. The decoding process was done in the same way as for SAD, as described in detail in Sect. 3.4.

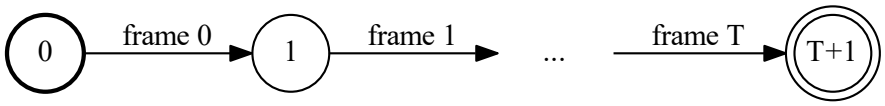

Figure 4.1: A transducer modeling the input signal for SCP detection.

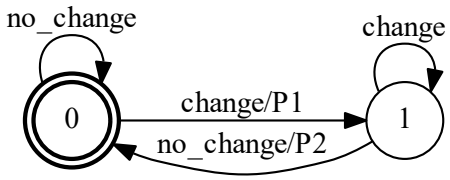

Figure 4.2: A transducer representing the transduction model for SCP detection.

The results are presented in the second row of Table 4.1. They show that the decoder was capable of operating in real time with an RTF of 0.022. This value, combined with the latency of 2.4 seconds, allowed it to be seamlessly used in an online environment. Although the achieved results provided a decent starting point, the precision was particularly weak and overshadowed by LIUM toolkit (i.e., 59.4% vs. 89.9%). Therefore, the next goal was to improve the quality of the SCP detection.

4.5 Enhanced Training Dataset

After thoroughly evaluating the results obtained so far, two types of errors were the most prominent. The first one was represented by change points omitted due to the quick artificial transitions between speakers (e.g., director cuts in broadcast news) while the second type resulted in change points falsely detected because of a silence longer than 0.5 seconds in speaker-homogeneous segments (caused by deep breaths or hesitation). As a solution to the first issue, 10 hours of recordings were prepared by artificially joining utterances of two different speakers. In total, 14,340 change points with a uniform distribution between all transition types (female-female, female-male,

[1]https://pytorch.org/

male-female, and male-male) were thus added to the training dataset. To reduce the latter type of errors, another 10 hours of training data were prepared. This data contains speaker-homogeneous segments with frequent occurrences of long silences.

The results gathered in the third row of Table 4.1 show that the use of enhanced training dataset led to significant improvement in all of the evaluation metrics observed. For example, the F-measure value got boosted up from 61.4% to 68.8%, while $\delta_{2/3}$ was enhanced to 0.21 seconds. Additionally, the average latency was slightly reduced, namely, from 2.4 seconds to 2.3 seconds.

4.6 Acoustic Features

In the next set of experiments, several feature extraction techniques were explored. In addition to the 39-dimensional MFCCs, 13-dimensional MFCCs with Δ and $\Delta\Delta$ coefficients (i.e., a 39-dimensional feature vector as well), and 39-dimensional bottleneck features were also utilized. As suggested, e.g., in [104–106], BTN features were extracted from DNN trained to discriminate physical states (senones) of a Czech tied-state triphone acoustic model. This deep extractor was trained on 270 hours of clean speech recordings of the Czech language. The hyper-parameters were set as follows: 5 hidden layers (the third one being the bottleneck layer), 1024 neurons per hidden layer (39 for the bottleneck layer), ReLU activation function (sigmoid for the bottleneck layer), mini-batches size of 1,024, 0.08 learning rate, and 50 epochs. 39-dimensional FBCs were used as input features, and the input feature vector was formed by concatenating 5 previous frames, the current frame, and 5 following frames. Local normalization within a one-second window was applied. More detailed information about the extractor and its performance in spoken language identification can be found in [107].

The results obtained are shown in Table 4.2. They show that the BTN features yielded significantly worse results in all of the observed metrics (e.g., the F-measure value dropped from 68.8% to 56.7%) and that they are more suitable for the tasks of language and speaker identification. On the contrary, the MFCCs with the Δ and $\Delta\Delta$ coefficients outperformed the originally chosen MFCC configuration. Both the quality and real-time performance of SCP detection improved (e.g., the latency was reduced from 2.3 seconds to 1.9 seconds because the decoder was able to make the final decisions more rapidly). A likely reason is additional information provided by the Δ and $\Delta\Delta$ coefficients.

Table 4.2: Results of the experiment exploring various feature extraction techniques.

features	P	R	F [%]	$\delta_{2/3}[s]$	RTF	L $[s]$
MFCCs	67.0	70.7	68.8	0.21	**0.022**	2.3
MFCCs + Δ + $\Delta\Delta$	**72.8**	**74.7**	**73.7**	**0.19**	0.024	**1.9**
BTNs	53.7	60.1	56.7	0.26	0.070	2.9

4.7 Convolutional Neural Networks

In the next step, more complex neural network architecture, CNN, was investigated. This architecture was employed for its feature representation and modeling capabilities. The utilized CNN was composed of two convolutional and two fully connected layers. The inputs consisted of 201 feature maps (i.e., 2-second context windows) in size of 39×1. The first convolutional layer was comprised of 105 feature maps at a size of 39×1, followed by a 3:1 max-pooling layer; the second one had 157 feature maps at a size of 13×1. The rest of the hyper-parameters remained unchanged.

The results are summarized in the fifth row of Table 4.1. The utilization of the CNNs yielded an overall improvement in all quality metrics (e.g., the F-measure value increased from 73.7% to 78.6%). The latency remained constant while the deterioration in RTF could be considered negligible (i.e., it is still significantly smaller than 1). For these reasons, CNNs were thus utilized for all follow-up experiments.

4.8 Context Window Size

The following experiments focused on the size of the input feature window. This additional context should result in a higher quality of the SCP detection at the cost of worse latency. Initially, a 2-second window had been chosen. In this experimental evaluation, the sizes ranging from 1 second up to 4 seconds were explored.

The results are in Table 4.3. As expected, the performance (i.e., F-measure and $\delta_{2/3}$) was further improved with the additional context (e.g., up to F-measure of 81.7%). On the contrary, the latency of the system was worsened with more context information by up to 2 seconds. The RTF remained relatively constant.

Table 4.3: Results exploring the influence of the context size on SCP detection.

context [s] (frames)	P	R	F [%]	$\delta_{2/3}[s]$	RTF	L [s]
1 (50-1-50)	71.3	69.4	70.3	0.21	**0.053**	**1.4**
1.5 (75-1-75)	71.0	72.8	71.9	**0.14**	**0.053**	1.7
2 (100-1-100)	79.3	77.8	78.6	0.17	0.054	1.9
2.5 (125-1-125)	80.3	81.8	81.1	0.17	0.054	2.3
3 (150-1-150)	80.0	**83.1**	81.5	0.17	0.054	2.6
3.5 (175-1-175)	**80.5**	82.6	81.5	0.16	0.055	3.1
4 (200-1-200)	80.4	**83.1**	**81.7**	0.16	0.055	3.5

4.9 WFST with a Forced Length of Transition

In the next experiments, the aim was to improve the results by introducing WFST with a forced transition model. This model was designed to reflect the annotation

style of the training data. As stated in Sect. 4.2, a 1-second window around the actual change point was labeled as speaker transition frames. However, during the decoding, the real duration of the transition between two speakers varied greatly.

Therefore, in this experiment, the duration of the transition was forced to be exactly 1 second at first. For this purpose, the transduction model was modified (see in Fig. 4.3) to correspond to the duration of the forced transition: it consists of two main states (0 and 1) and 98 transition states (shown as …).

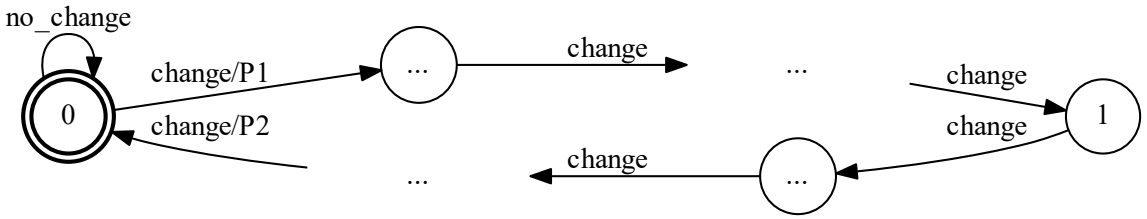

Figure 4.3: A transducer representing the transduction model with the forced transition for SCP detection.

This scheme works as follows: when a speaker change occurs, the decoder moves frame by frame from state 0 through half of the transition states to state 1. Here, a new change point label is provided, and the decoder moves backward to state 0, where it waits until the next change occurs. Note that, during this process, the penalty factors P1 and P2 (tuned on the development set) are in place as well.

The results are summarized in Table 4.4. First, a CNN with a context size of 2.5 seconds was used. Next, not only the forced length of the transition at 1 second but also several other values in a range from 0.5 up to 2 seconds were evaluated. The results show two contradictory trends: the quality of detection increased with the additional duration, while the RTF and latency values were worsened. Therefore, the optimal value of the duration strongly depends on the target application.

Table 4.4: Results studying varied durations of forced transitions in the WFST.

forced duration [s]	P	R	F [%]	$\delta_{2/3}[s]$	RTF	L [s]
0.5	77.2	75.2	76.2	**0.13**	**0.057**	**2.2**
1	82.7	**81.8**	82.2	0.17	0.065	2.9
1.5	83.5	81.5	82.5	0.16	0.072	3.7
2	**84.2**	81.5	**82.8**	0.17	0.079	4.5

For online application, the primary limiting factor is latency. In this environment, with the forced length of 1 second and total latency below 3 seconds, the proposed approach still allows for performing SCP detection with an accuracy level approaching the offline reference system (see the penultimate row of Table 4.1). As such, the online approach is ready to be integrated into the TVR monitoring system.

For offline application, where the latency and real-time processing are not an issue, it is possible to tune the proposed SCP detection approach to improve the

achieved results even further. For instance, a system based on CNN, the context window size of 3 seconds, and WFST with a forced length of 2 seconds yielded an F-measure value of 85.6% and a $\delta_{2/3}$ value of 0.18 seconds (with the latency at 4.8 seconds). These results are available for comparison in the last row of Table 4.1.

4.10 Evaluation on Whole COST278 Database

Until now, all of the conducted experiments were evaluated only on the Czech test subset of the COST278 [100, 101] database. In this experiment, both the proposed approach (tuned for online use) and the reference system were employed for SCP detection on the whole test dataset. The proposed approach was also trained only on the COST278 training data. In summary, it utilized MFCCs with the Δ and $\Delta\Delta$ coefficients, the CNN instead of the feed-forward DNN, an extended context size (2.5 seconds), and the WFST-based decoder with a 1-second forced transition. As suggested, the evaluation was done on all 11 languages of the test dataset, and the results were compared with the LIUM toolkit. The goal was to see if the proposed single-pass approach (without clustering) can compete with an offline reference tool.

The results show that both approaches perform on a relatively similar level. LIUM toolkit yielded an F-measure value of 73.5% and a $\delta_{2/3}$ value of 0.21 seconds, while the proposed approach scored an F-measure value of 73.1% and a $\delta_{2/3}$ value of 0.15 seconds, with the latency at 2.9 seconds. Figure 4.4 depicts the detailed results for all COST278 languages. The easiest ones were four closely related Slavic languages – Czech, Slovenian, Croatian and Slovak. Basque and Spanish for the LIUM toolkit and Belgian Dutch and Basque for the proposed SCP detection approach were the most difficult instances.

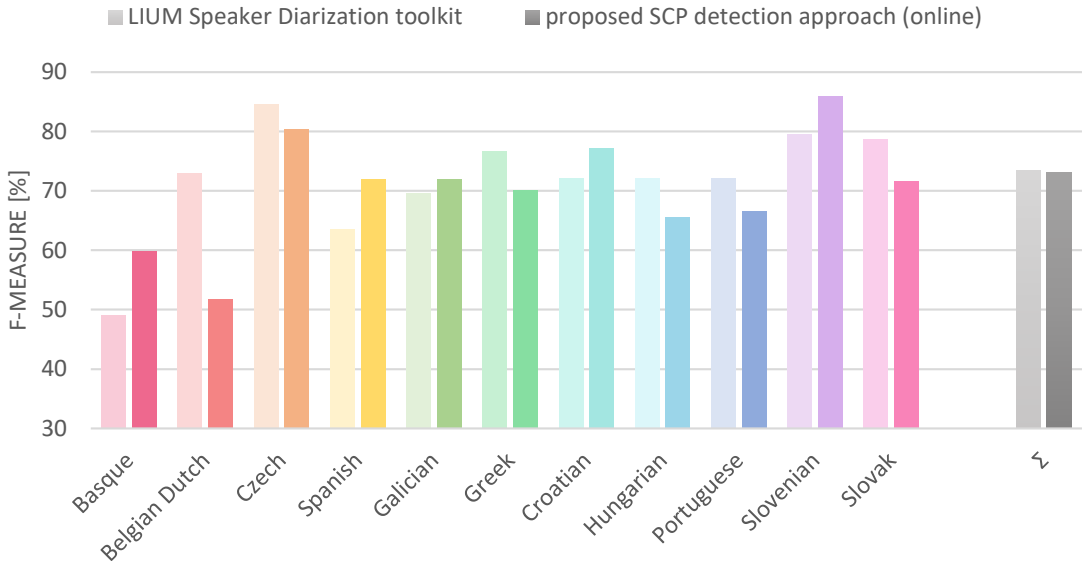

Figure 4.4: A comparison of the proposed SCP detection approach (tuned for online use) with the reference system on the whole COST278 database. Lighter columns mark the reference system while the darker ones indicate the proposed approach.

5 Conclusions

Within the scope of the thesis, the tasks of speech activity detection and speaker change point detection with the focus on modern technologies and their application in an online monitoring system as speech preprocessors have been explored. A novel approach has been proposed for speech activity detection as well as for speaker change point detection. The thesis closely follows and describes the development of both of these approaches from the initial to the final stages. All the steps taken are discussed in detail and backed up by a diverse set of experiments. Ultimately, both of these approaches have been designed to be integrated into the TVR monitoring system developed at SpeechLab@TUL in cooperation with the NanoTrix company, and they both support a crucial online mode.

Speech Activity Detection

The final proposed speech activity detection approach is based on two main components: a feed-forward deep neural network and a context-based weighted finite-state transducer. The first component, DNN, functions as a frame classifier (speech/non-speech and context states), while the latter component, WFST, is an online decoder which smooths the outputs of the classifier. The network is trained on log filter bank coefficients of artificially created data by mixing speech and non-speech recordings at various levels of SNR. The data has also been enriched by various noises. This design yields state-of-the-art results under low- and medium-noise conditions on the standardized QUT-NOISE-TIMIT dataset. Moreover, it also operates with a low real-time factor as well as low latency, which makes it a suitable option for online processing. An evaluation in a real speech transcription system has yielded a significant improvement in RTF as well as a slight boost in accuracy of the transcription.

The initial research introducing the main concept and a simple transduction model was presented in [88] at SIGMAP 2016 held in Lisbon. The improved and final context-based transduction model was introduced in [86] at ICASSP 2017 organized in New Orleans. Finally, an extended version detailing more experiments with QUT-NOISE-TIMIT corpus was published in [87].

Potential improvements could be focused on improving the latency even further. This could be achieved by, e.g., designing a different transduction model or employing diverse deep classifiers and fine-tuning their hyper-parameters. Additionally, more complex features could be crafted. Lastly, enrichment of training data by various broadcast noises could achieve more robust speech activity detection and yield even better speech/non-speech segmentation.

Speaker Change Point Detection

The final design of the proposed speaker change point detection approach is inspired by the proposed speech activity detection design. It consists of two main components: a convolutional neural network and a weighted finite-state transducer with a

forced length of transition. The convolutional neural network plays the role of a binary frame classifier (change point/no change point) while the weighted finite-state transducer is utilized as an online decoder smoothing the output of CNN. The decoder also enforces the duration of the transition from one speaker to another. The network is trained on TV/radio broadcast data complemented by artificial examples to reduce different types of errors. Safety collar frames are labeled around the actual change points to improve the performance of the system, and MFCCs with Δ and $\Delta\Delta$ are used as input features. On data taken from the COST278 database, the proposed approach achieves results approaching the offline multi-pass reference system (LIUM Speaker Diarization toolkit) while operating online with low latency.

The whole research explaining in detail the proposed speaker change point detection approach was presented in [99] at Interspeech 2019 conference in Graz.

The performance of the SCP detection approach could be further improved by implementing online clustering, which should diminish falsely predicted transitions between speakers. It is a common practice in the literature. An exploration of more robust features or different deep neural network architectures (e.g., time delay convolutional neural networks are gaining in popularity nowadays) could yield progress as well. Similarly to SAD, other transduction WFST models could be designed. Finally, additional varied training data could be collected to craft a more robust approach yielding even better results for diverse languages.

Summary of Research Contributions

Within the thesis, the following has been covered:

- an overview of the current state of the art in both speech activity detection and speaker change point detection with additional focus on existing toolkits;
- a detailed description of selected approaches to the SAD and SCP detection relevant to this work or focused on the online application;
- a detailed description of the design and development of the proposed SAD approach performing robust speech/non-speech detection;
- experimental tuning of the proposed SAD approach;
- an evaluation of the proposed SAD approach and its comparison with various SAD approaches on the standardized QUT-NOISE-TIMIT corpus;
- an evaluation of the proposed SAD approach in a real speech transcription system;
- a detailed description of the design and development of the proposed SCP detection approach performing speaker change point detection;
- experimental tuning of the proposed SCP detection approach;
- an evaluation of the proposed SCP detection approach and its comparison with a reference system on the standardized COST278 database;

- an evaluation of the online performance of both SAD and SCP detection approaches.

Summary of Practical Use Contributions

The main contribution of the thesis to the field of practical applications is the ability to integrate the proposed speech activity detection and speaker change point detection approaches into the TVR monitoring system developed at the author's lab in cooperation with the NanoTrix company.

The proposed SAD approach is now fully integrated into this TVR monitoring system. Last month, approximately 4,130 days (99,100 hours or 2.3 TB) of recordings were transcribed in the processing time of 1,333 days (32,000 hours). Considering the real-time factor of the speech transcriber being around one, the deployment of SAD (as a preprocessor) resulted in significantly saved processing time. Approximately two-thirds of the data was non-speech and thus omitted from the transcription. This was supplemented by a slight increase in accuracy of the transcriber as the non-speech parts were not transcribed into gibberish.

The proposed SCP detection approach is now ready to be integrated into this TVR monitoring system. Once done, it will be used to label speaker-homogeneous segments in multiple online broadcast streams (i.e., it will break the streams into smaller chunks, each containing only one speaker). By doing this, it will provide the transcribed data with additional information that could be further utilized and expanded upon. It will also form a stepping stone for further diarization functionality.

In general, both the SAD and SCP detection approaches can be used for any application that needs speech preprocessing, even the ones requiring online use.

Future Work

The fully implemented speech activity detection and speaker change point detection approaches are the first steps in the process of designing a speaker diarization system and successively speaker verification and identification systems and integrating them into a TVR monitoring system. In conjunction with SAD, the SCP detector produces an ever-growing amount of labels for speaker-homogeneous speech segments. These newly defined segments will be utilized for, e.g., language identification (the online version is already being worked on while the offline version was published in [107] at Interspeech 2018), gender, or emotion recognition. Their application to speaker-adaptive speech recognition is also planned in the future.

References

[1] G. Evangelopoulos and P. Maragos. "Speech Event Detection Using Multiband Modulation Energy". In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 685–688.

[2] B. Kotnik, Z. Kacic, and B. Horvat. "A Multiconditional Robust Front-End Feature Extraction with a Noise Reduction Procedure Based on Improved Spectral Subtraction Algorithm". In: *INTERSPEECH 2001 - Eurospeech, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 3-7, 2001*. ISCA, 2001, pp. 197–200.

[3] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan. "Noise Robust Voice Activity Detection Using Features Extracted from the Time-Domain Autocorrelation Function". In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, 2010, pp. 3118–3121.

[4] N. Ryant, M. Liberman, and J. Yuan. "Speech Activity Detection on YouTube Using Deep Neural Networks". In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 728–731.

[5] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah. "A Model Based Voice Activity Detector for Noisy Environments". In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 2297–2301.

[6] Y. Shao and Q. Lin. "Use of Pitch Continuity for Robust Speech Activity Detection". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5534–5538.

[7] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. V. Segbroeck, A. Potamianos, and S. Narayanan. "Multi-Band Long-Term Signal Variability Features for Robust Voice Activity Detection". In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 718–722.

[8] T. Kinnunen, A. Sholokhov, E. el Khoury, D. A. L. Thomsen, M. Sahidullah, and Z. Tan. "HAPPY Team Entry to NIST OpenSAD Challenge: A Fusion of Short-Term Unsupervised and Segment i-Vector Based Speech Activity Detectors". In: *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, California, USA, September 8-12, 2016*. ISCA, 2016, pp. 2992–2996.

[9] J. Ma. “Improving the Speech Activity Detection for the DARPA RATS Phase-3 Evaluation”. In: *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*. ISCA, 2014, pp. 1558–1562.

[10] L. Ferrer, M. Graciarena, and V. Mitra. “A phonetically aware system for speech activity detection”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5710–5714.

[11] J. W. Shin, J. Chang, and N. S. Kim. “Voice Activity Detection Based on Statistical Models and Machine Learning Approaches”. In: *Computer Speech & Language* 24.3 (2010), pp. 515–530.

[12] A. Misra. “Speech/Nonspeech Segmentation in Web Videos”. In: *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*. ISCA, 2012, pp. 1977–1980.

[13] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka. “Developing a Speech Activity Detection System for the DARPA RATS Program”. In: *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*. ISCA, 2012, pp. 1969–1972.

[14] H. Ghaemmaghami, D. Dean, S. Kalantari, S. Sridharan, and C. Fookes. “Complete-Linkage Clustering for Voice Activity Detection in Audio and Visual Speech”. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 2292–2296.

[15] L. Wang, C. Zhang, P. C. Woodland, M. J. F. Gales, P. Karanasou, P. Lanchantin, X. Liu, and Y. Qian. “Improved DNN-based segmentation for multi-genre broadcast audio”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5700–5704.

[16] I. Jang, C. Ahn, J. Seo, and Y. Jang. “Enhanced Feature Extraction for Speech Detection in Media Audio”. In: *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 479–483.

[17] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury. “The IBM Speech Activity Detection System for the DARPA RATS Program”. In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 3497–3501.

[18] R. Zazo, T. N. Sainath, G. Simko, and C. Parada. “Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection”. In: *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, California, USA, September 8-12, 2016*. ISCA, 2016, pp. 3668–3672.

[19] S. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals. “Temporal Modeling Using Dilated Convolution and Gating for Voice-Activity-Detection”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5549–5553.

[20] T. Hughes and K. Mierle. “Recurrent Neural Networks for Voice Activity Detection”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, British Columbia, Canada, May 26-31, 2013*. IEEE, 2013, pp. 7378–7382.

[21] F. Eyben, F. Weninger, S. Squartini, and B. W. Schuller. “Real-Life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, British Columbia, Canada, May 26-31, 2013*. IEEE, 2013, pp. 483–487.

[22] D. Karakos, S. Novotney, L. Zhang, and R. M. Schwartz. “Model Adaptation and Active Learning in the BBN Speech Activity Detection System for the DARPA RATS Program”. In: *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, California, USA, September 8-12, 2016*. ISCA, 2016, pp. 3678–3682.

[23] Q. Wang, J. Du, X. Bao, Z. Wang, L. Dai, and C. Lee. “A Universal VAD Based on Jointly Trained Deep Neural Networks”. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 2282–2286.

[24] X. Zhang and D. Wang. “Boosted Deep Neural Networks and Multi-Resolution Cochleagram Features for Voice Activity Detection”. In: *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*. ISCA, 2014, pp. 1534–1538.

[25] S. Thomas, G. Saon, M. V. Segbroeck, and S. S. Narayanan. “Improvements to the IBM Speech Activity Detection System for the DARPA RATS Program”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 4500–4504.

[26] J. Kim and M. Hahn. “Voice Activity Detection Using an Adaptive Context Attention Model”. In: *IEEE Signal Processing Letters* 25.8 (2018), pp. 1181–1185.

[27] H. Chung, S. J. Lee, and Y. Lee. “Endpoint Detection Using Weighted Finite State Transducer”. In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 700–703.

[28] C. Gao, G. Saikumar, S. Khanwalkar, A. Herscovici, A. Kumar, A. Srivastava, and P. Natarajan. “Online Speech Activity Detection in Broadcast News”. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 2637–2640.

[29] B. Liu, B. Hoffmeister, and A. Rastrow. “Accurate Endpointing with Expected Pause Duration”. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 2912–2916.

[30] D. Cournapeau and T. Kawahara. “Evaluation of real-time voice activity detection based on high order statistics”. In: *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*. ISCA, 2007, pp. 2945–2948.

[31] D. Cournapeau, S. Watanabe, A. Nakamura, and T. Kawahara. “Using online model comparison in the Variational Bayes framework for online unsupervised Voice Activity Detection”. In: *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, Dallas, Texas, USA, March 14-19, 2010*. IEEE, 2010, pp. 4462–4465.

[32] M. H. Moattar and M. M. Homayounpour. “A Simple but Efficient Real-Time Voice Activity Detection Algorithm”. In: *17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, United Kingdom, August 24-28, 2009*. IEEE, 2009, pp. 2549–2553.

[33] I. Tashev and S. Mirsamadi. “DNN-based Causal Voice Activity Detector”. In: *Information Theory and Applications Workshop*. University of California – San Diego, Feb. 2016.

[34] J. Zelinka. “Deep Learning and Online Speech Activity Detection for Czech Radio Broadcasting”. In: *Text, Speech, and Dialogue - 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018*. Springer, 2018, pp. 428–435.

[35] J. H. Ko, J. Fromm, M. Philipose, I. Tashev, and S. Zarar. “Limiting Numerical Precision of Neural Networks to Achieve Real-Time Voice Activity Detection”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April 15-20, 2018*. IEEE, 2018, pp. 2236–2240.

[36] D. Wang, L. Lu, and H. Zhang. “Speech segmentation without speech recognition”. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003, Hong Kong, April 6-10, 2003*. IEEE, 2003, pp. 468–471.

[37] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland. “The Cambridge University March 2005 Speaker Diarisation System”. In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 2437–2440.

[38] S. Meignier, D. Moraru, C. Fredouille, J. Bonastre, and L. Besacier. “Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization”. In: *Computer Speech & Language* 20.2-3 (2006), pp. 303–330.

[39] L. Lu, H. Zhang, and H. Jiang. “Content Analysis for Audio Classification and Segmentation”. In: *IEEE Transactions Speech and Audio Processing* 10.7 (2002), pp. 504–516.

[40] B. Desplanques, K. Demuynck, and J. Martens. “Factor Analysis for Speaker Segmentation and Improved Speaker Diarization”. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 3081–3085.

[41] L. V. Neri, H. N. B. Pinheiro, T. I. Ren, G. D. C. Cavalcanti, and A. G. Adami. “Speaker Segmentation Using i-vector in Meetings Domain”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 5455–5459.

[42] K. Chen and A. Salman. “Learning Speaker-Specific Characteristics With a Deep Neural Architecture”. In: *IEEE Transactions Neural Networks* 22.11 (2011), pp. 1744–1756.

[43] A. Sarkar, S. Dasgupta, S. K. Naskar, and S. Bandyopadhyay. “Says Who? Deep Learning Models for Joint Speech Recognition, Segmentation and Diarization”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5229–5233.

[44] R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng. “Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 5420–5424.

[45] H. Bredin. “TristouNet: Triplet Loss for Speaker Turn Embedding”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 5430–5434.

[46] A. Jati and P. G. Georgiou. “Speaker2Vec: Unsupervised Learning and Adaptation of a Speaker Manifold Using Deep Neural Networks with an Evaluation on Speaker Segmentation”. In: *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 3567–3571.

[47] A. Zhang, Q. Wang, Z. Zhu, J. W. Paisley, and C. Wang. “Fully Supervised Speaker Diarization”. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 6301–6305.

[48] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur. “Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge”. In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*. ISCA, 2018, pp. 2808–2812.

[49] S. S. Chen and P. S. Gopalakrishnan. “Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion”. In: *DARPA Broadcast News Transcription and Understanding Workshop*. 1998, pp. 127–132.

[50] P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia. “On the use of the Bayesian information criterion in multiple speaker detection”. In: *INTERSPEECH 2001 - Eurospeech, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September 3-7, 2001*. ISCA, 2001, pp. 795–798.

[51] M. Cettolo, M. Vescovi, and R. Rizzi. “Evaluation of BIC-Based Algorithms for Audio Segmentation”. In: *Computer Speech & Language* 19.2 (2005), pp. 147–170.

[52] H. Gish, M. H. Siu, and R. Rohlicek. “Segregation of Speakers for Speech Recognition and Speaker Identification”. In: *1991 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1991, Toronto, Ontario, Canada, May 14-17, 1991*. IEEE, 1991, pp. 873–876.

[53] C. Barras, X. Zhu, S. Meignier, and J. Gauvain. “Multistage Speaker Diarization of Broadcast News”. In: *IEEE Transactions Audio, Speech & Language Processing* 14.5 (2006), pp. 1505–1512.

[54] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. “Automatic Segmentation, Classification and Clustering of Broadcast News Audio”. In: *DARPA Speech Recognition Workshop*. 1997, pp. 97–99.

[55] B. Fergani, M. Davy, and A. Houacine. “Speaker diarization using one-class support vector machines”. In: *Speech Communication* 50.5 (2008), pp. 355–365.

[56] S. Meignier, J. Bonastre, and S. Igounet. “E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing”. In: *2001: A Speaker Odyssey - The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001*. ISCA, 2001, pp. 175–180.

[57] A. S. Malegaonkar, A. M. Ariyaeeinia, and P. Sivakumaran. “Efficient Speaker Change Detection Using Adapted Gaussian Mixture Models”. In: *IEEE Transactions Audio, Speech & Language Processing* 15.6 (2007), pp. 1859–1869.

[58] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. “Stream-Based Speaker Segmentation Using Speaker Factors and Eigenvoices”. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, Las Vegas, Nevada, USA, March 30 - April 4, 2008*. IEEE, 2008, pp. 4133–4136.

[59] V. Gupta. “Speaker Change Point Detection Using Deep Neural Nets”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 4420–4424.

[60] M. Hruz and M. Kunesova. “Convolutional Neural Network in the Task of Speaker Change Detection”. In: *Speech and Computer - 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016*. Springer, 2016, pp. 191–198.

[61] M. Hruz and Z. Zajic. “Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 4945–4949.

[62] M. India, J. A. R. Fonollosa, and J. Hernando. “LSTM Neural Network-Based Speaker Segmentation Using Acoustic and Language Modelling”. In: *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 2834–2838.

[63] R. Yin, H. Bredin, and C. Barras. “Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks”. In: *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 3827–3831.

[64] M. Hruz and M. Hlavac. “LSTM Neural Network for Speaker Change Detection in Telephone Conversations”. In: *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018*. Springer, 2018, pp. 226–233.

[65] L. Lu and H. Zhang. “Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis”. In: *10th ACM International Conference on Multimedia 2002, Juan les Pins, France, December 1-6, 2002*. ACM, 2002, pp. 602–610.

[66] L. Lu and H. Zhang. “Unsupervised Speaker Segmentation and Tracking in Real-Time Audio Content Analysis”. In: *Multimedia Systems* 10.4 (2005), pp. 332–343.

[67] M. Kotti, L. P. M. Martins, E. Benetos, J. S. Cardoso, and C. Kotropoulos. “Automatic Speaker Segmentation using Multiple Features and Distance Measures: A Comparison of Three Approaches”. In: *2006 IEEE International Conference on Multimedia and Expo, ICME 2006, Toronto, Ontario, Canada, July 9-12, 2006*. IEEE, 2006, pp. 1101–1104.

[68] M. Grasic, M. Kos, and Z. Kacic. “Online speaker segmentation and clustering using cross-likelihood ratio calculation with reference criterion selection”. In: *IET Signal Processing* 4.6 (2010), pp. 673–685.

[69] X. Anguera. “Xbic: Real-time cross probabilities measure for speaker segmentation”. In: *ICSI* (2005), pp. 1–8.

[70] J. Ajmera, I. McCowan, and H. Bourlard. “Robust speaker change detection”. In: *IEEE Signal Processing Letters* 11.8 (2004), pp. 649–651.

[71] K. Markov and S. Nakamura. “Never-ending learning system for on-line speaker diarization”. In: *2007 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007*. IEEE, 2007, pp. 699–704.

[72] J. T. Geiger, F. Wallhoff, and G. Rigoll. “GMM-UBM Based Open-Set Online Speaker Diarization”. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, 2010, pp. 2330–2333.

[73] T. Wu, L. Lu, K. Chen, and H. Zhang. “Universal Background Models for Real-time Speaker Change Detection”. In: *9th International Conference on Multi-Media Modeling, MMM 2003, Taiwan, January 7-10, 2003*. IEEE, 2003, pp. 135–149.

[74] D. Dimitriadis and P. Fousek. “Developing On-Line Speaker Diarization System”. In: *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 2739–2743.

[75] W. Zhu and J. W. Pelecanos. “Online Speaker Diarization Using Adapted i-vector Transforms”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5045–5049.

[76] Z. Ge, A. N. Iyer, S. Cheluvaraja, and A. Ganapathiraju. “Speaker Change Detection Using Features through a Neural Network Speaker Classifier”. In: *2017 Intelligent Systems Conference (IntelliSys), 2017, London, United Kingdom, September 7-8, 2017*. IEEE, 2017, pp. 1111–1116.

[77] M. Kunesova, Z. Zajic, and V. Radova. “Experiments with Segmentation in an Online Speaker Diarization System”. In: *Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017*. Springer, 2017, pp. 429–437.

[78] G. E. Hinton, S. Osindero, and Y. W. Teh. “A Fast Learning Algorithm for Deep Belief Nets”. In: *Neural Computation* 18.7 (2006), pp. 1527–1554.

[79] G. E. Dahl, D. Yu, L. Deng, and A. Acero. “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”. In: *IEEE Transactions Audio, Speech & Language Processing* 20.1 (2012), pp. 30–42.

[80] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[81] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu. “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin”. In: *CoRR* abs/1512.02595 (2015).

[82] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Nevada, USA, December 3-6, 2012*. Curran Associates, Inc., 2012, pp. 1106–1114.

[83] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, Nevada, USA, June 27-30, 2016*. IEEE, 2016, pp. 770–778.

[84] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Nevada, USA, December 5-8, 2013*. Curran Associates, Inc., 2013, pp. 3111–3119.

[85] I. Sutskever, O. Vinyals, and Q. V. Le. “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems 27: 28th Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, December 8-13, 2014*. Curran Associates, Inc., 2014, pp. 3104–3112.

[86] L. Mateju, P. Cerva, J. Zdansky, and J. Malek. “Speech Activity Detection in Online Broadcast Transcription Using Deep Neural Networks and Weighted Finite State Transducers”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 5460–5464.

[87] L. Mateju, P. Cerva, and J. Zdansky. “Investigation into the Use of WFSTs and DNNs for Speech Activity Detection in Broadcast Data Transcription”. In: *E-Business and Telecommunications - 13th International Joint Conference, ICETE 2016, Lisbon, Portugal, July 26-28, 2016, Revised Selected Papers*. Springer, 2017, pp. 341–358.

[88] L. Mateju, P. Cerva, and J. Zdansky. “Study on the Use of Deep Neural Networks for Speech Activity Detection in Broadcast Recordings”. In: *13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 5: SIGMAP, Lisbon, Portugal, July 26-28, 2016*. SciTePress, 2016, pp. 45–51.

[89] D. Dean, S. Sridharan, R. Vogt, and M. Mason. “The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms”. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, 2010, pp. 3110–3113.

[90] O. J. Rasanen, U. K. Laine, and T. Altosaar. “An Improved Speech Segmentation Quality Measure: the R-value”. In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. ISCA, 2009, pp. 1851–1854.

[91] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. “The DARPA Speech Recognition Research Database: Specifications and Status”. In: *Proceedings of DARPA Workshop on Speech Recognition*. 1986, pp. 93–99.

[92] S. Wisdom, G. Okopal, L. E. Atlas, and J. W. Pitton. “Voice Activity Detection Using Subband Noncircularity”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 4505–4509.

[93] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit. “ITU-T Recommendation G.729 Annex B: a Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications”. In: *IEEE Communications Magazine* 35.9 (Sept. 1997), pp. 64–73.

[94] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms". In: *ETSI ES 202 050 V1.1.5* (2007).

[95] J. Li, B. Liu, R. Wang, and L. Dai. "A Complexity Reduction of ETSI Advanced Front-End for DSR". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*. IEEE, 2004, pp. 61–64.

[96] J. Sohn, N. S. Kim, and W. Sung. "A Statistical Model-Based Voice Activity Detection". In: *IEEE Signal Processing Letters* 6.1 (1999), pp. 1–3.

[97] J. Ramirez, J. C. Segura, M. C. Benitez, A. de la Torre, and A. J. Rubio. "Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information". In: *Speech Communication* 42.3-4 (2004), pp. 271–287.

[98] L. Mateju, P. Cerva, and J. Zdansky. "Investigation into the Use of Deep Neural Networks for LVCSR of Czech". In: *2015 IEEE International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics, ECMSM, 2015, Liberec, Czech Republic, June 22-24, 2015*. IEEE, 2015, pp. 184–187.

[99] L. Mateju, P. Cerva, and J. Zdansky. "An Approach to Online Speaker Change Point Detection Using DNNs and WFSTs". In: *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, September 15-19, 2019*. ISCA, 2019, pp. 649–653.

[100] A. Vandecatseye, J. Martens, J. P. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez-Tirado, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, and C. Alexandris. "The COST278 Pan-European Broadcast News Database". In: *Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 26-28, 2004*. ELRA, 2004, pp. 873–876.

[101] J. Zibert, F. Mihelic, J. Martens, H. Meinedo, J. P. Neto, L. D. Fernandez, C. Garcia-Mateo, P. David, J. Zdansky, M. Pleva, A. Cizmar, A. Zgank, Z. Kacic, C. Teleki, and K. Vicsi. "The COST278 broadcast news segmentation and speaker clustering evaluation - overview, methodology, systems, results". In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, 2005, pp. 629–632.

[102] S. Meignier and T. Merlin. "LIUM SpkDiarization: an Open Source Toolkit for Diarization". In: *CMU SPUD Workshop, Dallas, Texas, USA, March 13, 2010*. 2010.

[103] M. Rouvier, G. Dupuy, P. Gay, E. el Khoury, T. Merlin, and S. Meignier. "An Open-Source State-of-the-Art Toolbox for Broadcast News Diarization". In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 1477–1481.

[104] F. Richardson, D. A. Reynolds, and N. Dehak. "A Unified Deep Neural Network for Speaker and Language Recognition". In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 1146–1150.

[105] M. McLaren, L. Ferrer, and A. Lawson. "Exploring the Role of Phonetic Bottleneck Features for Speaker and Language Recognition". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5575–5579.

[106] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer. "Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition". In: *IEEE/ACM Transactions Audio, Speech & Language Processing* 24.1 (2016), pp. 105–116.

[107] L. Mateju, P. Cerva, J. Zdansky, and R. Safarik. "Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal". In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*. ISCA, 2018, pp. 1803–1807.

[108] X. Zhang and J. Wu. "Deep Belief Networks Based Voice Activity Detection". In: *IEEE Transactions Audio, Speech & Language Processing* 21.4 (2013), pp. 697–710.

[109] M. V. Segbroeck, A. Tsiartas, and S. Narayanan. "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice". In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 704–708.

[110] M. Graciarena, L. Ferrer, and V. Mitra. "The SRI System for the NIST OpenSAD 2015 Speech Activity Detection Evaluation". In: *INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, California, USA, September 8-12, 2016*. ISCA, 2016, pp. 3673–3677.

[111] Y. Obuchi. "Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5715–5719.

[112] D. Snyder, G. Chen, and D. Povey. "MUSAN: A Music, Speech, and Noise Corpus". In: *CoRR* abs/1510.08484 (2015).

[113] S. Chaudhuri, J. Roth, D. P. W. Ellis, A. C. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. W. Wilson, and Z. Xi. "AVA-Speech: A Densely Labeled Dataset of Speech Activity in Movies". In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication*

Association, Hyderabad, India, September 2-6, 2018. ISCA, 2018, pp. 1239–1243.

[114] P. Delacourt and C. Wellekens. "DISTBIC: A speaker-based segmentation for audio data indexing". In: *Speech Communication* 32.1-2 (2000), pp. 111–126.

[115] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J. Bonastre. "The ELISA Consortium Approaches in Broadcast News Speaker Segmentation During the NIST 2003 Rich Transcription Evaluation". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*. IEEE, 2004, pp. 373–376.

[116] R. Stern. "Specifications of the 1996 Hub-4 Broadcast News Evaluation". In: *DARPA Speech Recognition Workshop*. 1997.

[117] S. Galliano, E. Geoffrois, G. Gravier, J. Bonastre, D. Mostefa, and K. Choukri. "Corpus Description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News". In: *Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*. ELRA, 2006, pp. 139–142.

[118] O. Galibert, J. Leixa, G. Adda, K. Choukri, and G. Gravier. "The ETAPE Speech Processing Evaluation". In: *Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*. ELRA, 2014, pp. 3995–3999.

[119] O. Galibert and J. Kahn. "The First Official REPERE Evaluation". In: *First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22-23, 2013*. CEUR-WS.org, 2013, pp. 43–48.

[120] J. Bonastre, F. Wils, and S. Meignier. "ALIZE, a Free Toolkit for Speaker Recognition". In: *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2005, Philadelphia, Pennsylvania, USA, March 18-23, 2005*. IEEE, 2005, pp. 737–740.

[121] A. Larcher, J. Bonastre, B. G. B. Fauve, K. Lee, C. Levy, H. Li, J. S. D. Mason, and J. Parfait. "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition". In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. ISCA, 2013, pp. 2768–2772.

[122] S. Bozonnet, N. W. D. Evans, and C. Fredouille. "The Lia-Eurecom RT'09 Speaker Diarization System: Enhancements in Speaker Modelling and Cluster Purification". In: *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, Dallas, Texas, USA, March 14-19, 2010*. ISCA, 2010, pp. 4958–4961.

[123] E. el Khoury, L. E. Shafey, and S. Marcel. "Spear: An open source toolbox for speaker recognition based on Bob". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 2014, pp. 1655–1659.

[124] D. Vijayasenan and F. Valente. “DiarTk : An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings”. In: *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*. ISCA, 2012, pp. 2170–2173.

[125] R. Yin, H. Bredin, and C. Barras. “Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization”. In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*. ISCA, 2018, pp. 1393–1397.

[126] P. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier. “S4D: Speaker Diarization Toolkit in Python”. In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*. ISCA, 2018, pp. 1368–1372.

[127] A. Larcher, K. Lee, and S. Meignier. “An extensible speaker identification sidekit in Python”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5095–5099.

[128] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. “X-Vectors: Robust DNN Embeddings for Speaker Recognition”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5329–5333.

[129] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. “The Kaldi Speech Recognition Toolkit”. In: *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, Hawaii, USA, December 11-15, 2011*. IEEE, 2011, pp. 1–4.

[130] “Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)”. In: *ITU-T Recommendation G.729* (1996), pp. 1–152.

[131] R. Salami, C. Laflamme, B. Bessette, and J. P. Adoul. “ITU-T G.729 Annex A: Reduced Complexity 8 kb/s CS-ACELP Codec for Digital Simultaneous Voice and Data”. In: *IEEE Communications Magazine* 35.9 (Sept. 1997), pp. 56–63.

[132] J. Sohn and W. Sung. “A voice activity detector employing soft decision based noise spectrum adaptation”. In: *1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998, Seattle, Washington, USA, May 12-15, 1998*. IEEE, 1998, pp. 365–368.

[133] P. Mermelstein. “Distance Measures for Speech Recognition: Psychological and Instrumental”. In: *Pattern Recognition and Artificial Intelligence*. Academic Press, 1976, pp. 374–388.

[134] J. Wu and X. Zhang. “An efficient voice activity detection algorithm by combining statistical model and energy detection”. In: *EURASIP Journal on Advances in Signal Processing* 2011 (2011), pp. 18–27.

[135] J. Macqueen. “Some methods for classification and analysis of multivariate observations”. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. 1967, pp. 281–297.

[136] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society, Series B* 39.1 (1977), pp. 1–38.

[137] S. M. Strassel, A. Morris, J. G. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. “Creating HAVIC: Heterogeneous Audio Visual Internet Collection”. In: *Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*. ELRA, 2012, pp. 2573–2577.

[138] L. R. Rabiner. “Readings in Speech Recognition”. In: Morgan Kaufmann Publishers Inc., 1990. Chap. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296.

[139] D. Liu and F. Kubala. “Fast speaker change detection for broadcast news transcription and indexing”. In: *INTERSPEECH 1999 - Eurospeech, 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 5-9, 1999*. ISCA, 1999.

[140] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, Massachusetts, USA, June 28 - July 1, 2001*. 2001, pp. 282–289.

[141] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda. “Voice activity detection based on conditional random fields using multiple features”. In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, 2010, pp. 2086–2089.

[142] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999. ISBN: 978-0-387-98793-4.

[143] S. Galliano, G. Gravier, and L. Chaubard. “The ester 2 evaluation campaign for the rich transcription of French radio broadcasts”. In: *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. ISCA, 2009, pp. 2583–2586.

[144] D. Wang, R. Vogt, M. Mason, and S. Sridharan. “Automatic audio segmentation using the Generalized Likelihood Ratio”. In: *2nd International Conference on Signal Processing and Communication Systems (ICSPCS), Gold Coast, Australia, December 15-17, 2008*. IEEE, 2008, pp. 1–5.

[145] L. Lu, H. Jiang, and H. Zhang. “A robust audio classification and segmentation method”. In: *9th ACM International Conference on Multimedia 2001, Ottawa, Ontario, Canada, September 30 - October 5, 2001*. ACM, 2001, pp. 203–211.

[146] L. Lu, H. Zhang, and S. Z. Li. “Content-based audio classification and segmentation by using support vector machines”. In: *Multimedia Systems* 8.6 (2003), pp. 482–492.

[147] F. Itakura. “Line spectrum representation of linear predictor coefficients of speech signals”. In: *The Journal of the Acoustical Society of America* 57.S1 (1975), S35.

[148] S. Kullback and R. A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86.

[149] J. P. Campbell. “Speaker recognition: a tutorial”. In: *Proceedings of the IEEE* 85.9 (Sept. 1997), pp. 1437–1462.

[150] A. Tritschler and R. A. Gopinath. “Improved speaker segmentation and segments clustering using the bayesian information criterion”. In: *INTERSPEECH 1999 - Eurospeech, 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 5-9, 1999*. ISCA, 1999, pp. 679–682.

[151] A. S. Malegaonkar, A. M. Ariyaeeinia, P. Sivakumaran, and J. Fortuna. “Unsupervised speaker change detection using probabilistic pattern matching”. In: *IEEE Signal Processing Letters* 13.8 (2006), pp. 509–512.

[152] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE Transactions Audio, Speech & Language Processing* 19.4 (2011), pp. 788–798.

[153] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason. “i-vector Based Speaker Recognition on Short Utterances”. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 2341–2344.

[154] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass. “Exploiting Intra-Conversation Variability for Speaker Diarization”. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 945–948.

[155] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. “Language Recognition via i-vectors and Dimensionality Reduction”. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 857–860.

[156] D. M. Gonzalez, O. Plchot, L. Burget, O. Glembek, and P. Matejka. “Language Recognition in iVectors Space”. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*. ISCA, 2011, pp. 861–864.

[157] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. “Joint Factor Analysis Versus Eigenchannels in Speaker Recognition”. In: *IEEE Transactions Audio, Speech & Language Processing* 15.4 (2007), pp. 1435–1447.

[158] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. “The AMI Meeting Corpus: A Pre-announcement”. In: *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, United Kingdom, July 11-13, 2005, Revised Selected Papers*. Springer, 2005, pp. 28–39.

[159] B. Zhou and J. H. L. Hansen. “Efficient audio stream segmentation via the combined T^2 statistic and Bayesian information criterion”. In: *IEEE Transactions Speech and Audio Processing* 13.4 (2005), pp. 467–474.

[160] E. Variani, X. Lei, E. McDermott, J. Gonzalez-Dominguez, and I. Lopez-Moreno. “Deep neural networks for small footprint text-dependent speaker verification”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 2014, pp. 4052–4056.

[161] C. Cieri, D. Miller, and K. Walker. “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text”. In: *Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, May 26-28, 2004*. ELRA, 2004.

[162] L. Li, D. Wang, Z. Zhang, and T. F. Zheng. “Deep Speaker Vectors for Semi Text-independent Speaker Verification”. In: *CoRR* abs/1505.06427 (2015).

[163] Z. Zajic, M. Kunesova, and V. Radova. “Investigation of Segmentation in i-Vector Based Speaker Diarization of Telephone Speech”. In: *Speech and Computer - 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016*. Springer, 2016, pp. 411–418.

[164] A. H. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang. “Phoneme recognition using time-delay neural networks”. In: *IEEE Transactions Acoustics, Speech, and Signal Processing* 37.3 (1989), pp. 328–339.

[165] R. Kneser and H. Ney. “Improved Backing-off for M-gram Language Modeling”. In: *1995 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995, Detroit, Michigan, USA, May 8-12, 1995*. IEEE, 1995, pp. 181–184.

Author's Publications

2019:

1. L. Mateju, Z. Callejas, D. Griol, J. M. Molina, and A. Sanchis. "An Empirical Assessment of Deep Learning Approaches to Task-Oriented Dialog Management". Accepted to: *Neurocomputing (Q1)*. 2019.

2. L. Mateju, P. Cerva, and J. Zdansky. "An Approach to Online Speaker Change Point Detection Using DNNs and WFSTs". In: *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, September 15-19, 2019*. ISCA, 2019, pp. 649–653.

2018:

3. R. Safarik, L. Mateju, and L. Weingartova. "The Influence of Errors in Phonetic Annotations on Performance of Speech Recognition System". In: *Text, Speech, and Dialogue - 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018*. Springer, 2018, pp. 419–427.

4. L. Mateju, P. Cerva, J. Zdansky, and R. Safarik. "Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal". In: *INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, September 2-6, 2018*. ISCA, 2018, pp. 1803–1807.

5. R. Safarik and L. Mateju. "Automatic Development of ASR System for an Under-Resourced Language". In: *41st International Conference on Telecommunications and Signal Processing, TSP 2018, Athens, Greece, July 4-6, 2018*. IEEE, 2018, pp. 100–103.

2017:

6. L. Mateju, P. Cerva, and J. Zdansky. "Investigation into the Use of WFSTs and DNNs for Speech Activity Detection in Broadcast Data Transcription". In: *E-Business and Telecommunications - 13th International Joint Conference, ICETE 2016, Lisbon, Portugal, July 26-28, 2016, Revised Selected Papers*. Springer, 2017, pp. 341–358.

7. R. Safarik and L. Mateju. "The Impact of Inaccurate Phonetic Annotations on Speech Recognition Performance". In: *Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017*. Springer, 2017, pp. 402–410.

8. L. Mateju, P. Cerva, J. Zdansky, and J. Malek. "Speech Activity Detection in Online Broadcast Transcription Using Deep Neural Networks and Weighted Finite State Transducers". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, Louisiana, USA, March 5-9, 2017*. IEEE, 2017, pp. 5460–5464.

2016:

9. M. Bohac, L. Mateju, M. Rott, and R. Safarik. "Automatic Syllabification and Syllable Timing of Automatically Recognized Speech - for Czech". In: *Text, Speech, and Dialogue - 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016*. Springer, 2016, pp. 540–547.

10. L. Mateju, P. Cerva, and J. Zdansky. "Study on the Use of Deep Neural Networks for Speech Activity Detection in Broadcast Recordings". In: *13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 5: SIGMAP, Lisbon, Portugal, July 26-28, 2016*. SciTePress, 2016, pp. 45–51.

11. R. Safarik and L. Mateju. "Impact of Phonetic Annotation Precision on Automatic Speech Recognition Systems". In: *39th International Conference on Telecommunications and Signal Processing, TSP 2016, Vienna, Austria, June 27-29, 2016*. IEEE, 2016, pp. 311–314.

2015:

12. L. Mateju, P. Cerva, and J. Zdansky. "Investigation into the Use of Deep Neural Networks for LVCSR of Czech". In: *2015 IEEE International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics, ECMSM, 2015, Liberec, Czech Republic, June 22-24, 2015*. IEEE, 2015, pp. 184–187.