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

Speaker Change Point Detection Based on Deep Neural Networks for Online Streams

Thesis Statement

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

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

Liberec 2017

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Detekce změny mluvčího s využitím hlubokých neuronových sítí v online vysílání

Teze disertační práce

Studijní program: P2612 – Elektrotechnika a informatika
Studijní obor: 2612V045 – Technická kybernetika

Autor práce: **Ing. Lukáš Matějů**
Vedoucí práce: Ing. Petr Červa, Ph.D.

Liberec 2017

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Ing. Petr Červa, Ph.D., for his flawless guidance and continuous support of my Ph.D. study. Besides my supervisor, I am also grateful to Ing. Jindřich Žďánský, Ph.D. for his thorough insights and endless help. This work would not be possible without them. Last but not the least, I'd like to thank my friends & family and especially Dany for being there for me.

Abstract

This thesis statement is focused on a task of speaker change point detection for online broadcast streams. It summarizes the state-of-the-art approaches, explains and sets the motivation and main goals for final thesis and presents the current status of the work. The proposed approach for speaker change point detection operates in two consecutive phases: speech activity detection and speaker segmentation. Both phases are heavily discussed within this work.

The proposed speech activity detection approach utilizes deep neural networks as classifier (trained on artificially mixed speech and non-speech signals at desired level of SNR), and context-based weighted finite state transducers as online decoder (to smooth the outputs of DNN). This approach yields state-of-the-art results on standardized QUT-NOISE-TIMIT dataset, and it is also capable of operating in online mode (i.e., with low latency & real-time factor). The proposed speech activity detection approach is now completed, published and fully integrated into TVR monitoring system developed at SpeechLab@TUL, where it saves a significant amount of processing time.

The speaker segmentation approach, similarly to the speech activity detection one, is based on deep neural networks (trained on a mixture of real & artificial broadcast data) and weighted finite state transducers with forced duration of transition (to determine the change points). The achieved results on standardized COST278 dataset are promising, but further research is still needed. A publication is planned for next year.

The final thesis will build on (speech activity detection) and further extend this work (speaker segmentation). The final proposed speaker change point detection approach will be then fully integrated into TVR monitoring system developed at SpeechLab@TUL, where it will be utilized to process a large amount of online broadcast streams.

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

Abstrakt

Teze disertační práce se věnují úloze detekce změny mluvčího v online vysílání. Shrnují současné poznání ve světě, představují motivaci a cíle disertační práce a zabývají se jejím současným stavem. Metoda pro detekci změny mluvčího, představená v této práci, pracuje ve dvou po sobě jdoucích krocích. Prvním krokem je detekce řeči, která je následována samotnou detekcí změny mluvčího. Oběma krokům je věnována vysoká pozornost.

Metoda pro detekci řeči je založená na hlubokých neuronových sítích, které slouží jako klasifikátor, a na vážených konečných stavových transducerech, které vyhlazují výstup ze sítě a zároveň plní roli online dekodéru. Takto navržená metoda dosahuje nejlepších výsledků na datech QUT-NOISE-TIMIT. Je také možné ji bezproblémově nasadit pro online zpracování (nízká latence). Metoda je považována za dokončenou, byla publikována na mezinárodních konferencích, a dnes je již plně integrována do monitorovacího systému TVR vyvíjeného Laboratoří počítačového zpracování řeči TUL, kde šetří velké množství výpočetního času.

Podobně jako metoda pro detekci řeči je i metoda pro detekci změny mluvčího založená na hlubokých neurnových sítích a na vážených konečných stavových transducerech (tentokrát s pevnou délkou přechodu mezi mluvčími). Metoda je stále vylepšována, jelikož dosažené výsledky na datech COST278 nejsou plně uspokojující. První publikace je plánována na příští rok.

Po dokončení se obě metody stanou základním kamenem disertační práce. Výsledná metoda bude následně začleněna do monitorovacího systému TVR vyvíjeného Laboratoří počítačového zpracování řeči TUL, kde bude využita pro zpracování velkého objemu online vysílání.

Klíčová slova: detekce řeči, detekce změny mluvčího, hluboké neuronové sítě, online zpracování, vážené konečné stavové transducery.

Contents

List of Figures

List of Tables

List of Abbreviations

BIC	Bayesian Information Criterion
CNN	Convolutional Neural Networks
DBN	Deep Belief Networks
DNN	Deep Neural Networks
FAR	False Alarm Rate
FER	Frame Error Rate
HMM	Hidden Markov Models
HTER	Half-Total Error Rate
GMM	Gaussian Mixture Models
MFCC	Mel-Frequency Cepstral Coefficients
MR	Miss Rate
SNR	Signal-to-Noise Ratio
SAD	Speech Activity Detection
SCP	Speaker Change Point Detection
SVM	Support Vector Machines
RNN	Recurrent Neural Networks
RTF	Real-Time Factor
VAD	Voice Activity Detection
WER	Word Error Rate
WFST	Weighted Finite State Transducers

Introduction

Nowadays, a huge amount of audio data is produced every day by television and radio streams as well as other sources. However, most of this data lacks any kind of labels, which would be useful for a wide range of speech processing applications. These labels can also carry time marks, which can be further utilized for audio searching, indexing or data retrieval. Speaker change point detection (often called speaker segmentation) is one of the tasks that can create such labels. It is a task of finding precise change points between two speakers in a recording. The output should be labels of speaker homogeneous segments. Speaker segmentation is usually done without any prior knowledge about the identity or even the number of speakers in recording (i.e., it is treated as a speaker independent task).

Speaker change point detection has many uses as a speech preprocessing task. In conjunction with speaker clustering, it forms a speaker diarization system. Speaker diarization focuses on answering the question "who spoke when?", and the research is driven by challenges held by National Institute of Standards and Technology (NIST). Speaker change point detection is also a vital part of speaker verification & identification systems. It can further be applied for language, gender or emotion detection as well. The extracted segments can also be used as training data for speaker adaptive approaches to speech recognition. Lastly, speaker segmentation can be employed for tasks such as audio indexing and retrieval, rich transcription, automatic speech transcription, movie analysis (dialog detection), speaker tracking, multi-speaker detection, etc.

The varied application of speaker change point detection makes it a popular research topic. Numerous research groups and research centers compete worldwide and propose different approaches in a pursue of state-of-the-art results. Sometimes, challenges are also being held (e.g. segmentation task by NIST) in an effort to push the field further. The popularity of this research topic can also be documented by large amounts of accepted papers to international conferences on signal and/or speech processing, such as Interspeech or International Conference on Acoustics, Speech and Signal Processing (ICASSP). With the recent boom in deep learning in mind, speaker segmentation attracts more and more researchers, and a lot of interesting work should get published in near future.

The remainder of this thesis statement is organized as follows. In Sect. 1, state-of-the-art approaches to speaker segmentation (and speech activity detection) are summarized as well as an overview of existing systems is presented. Section 2 explains the motivation behind this thesis and it also sets its main goals. The current progress of final thesis is in detail discussed in Sect. 3. Within this section, the proposed speech activity detection and speaker change point detection approaches are presented and their current status is explained as well. Finally, the work is concluded in Sect. 4.

1 State-of-the-Art

In literature, speaker change point detection is usually carried out in two consecutive phases. The first phase is speech activity detection, a task of segmenting speech events from non-speech ones. The extracted speech segments can be optionally further preprocessed by, e.g., gender or language identification. The latter phase is the speaker change point detection itself. During this phase, the final change points between speakers (in extracted speech segments) are found and labeled.

Thanks to the varied application and usefulness of speaker segmentation, there are several fully working systems, which are freely available for download. Some of them even come with pretrained models for immediate use.

1.1 Speech Activity Detection

Most of the existing speech activity detection approaches operate in two subsequent stages: feature extraction and speech/non-speech classification.

In the former phase, the classic approaches for feature extraction utilize energy (Evangelopoulos and Maragos, 2005), zero crossing rate (Kotnik, Kacic, and Horvat, 2001) or auto-correlation function (Ghaemmaghami, Baker, et al., 2010). The family of more complex features, which have also been successfully applied, includes MFCCs (Sriskandaraja et al., 2015; Ryant, Liberman, and Yuan, 2013), multi-resolution cochleagram features (X. Zhang and D. Wang, 2014), multi-band long-term signal variability features (Tsiartas et al., 2013) or channel bottleneck features (Ma, 2014). Note that in (X. Zhang and Wu, 2013) features based on the use of deep belief networks have also been proposed. In practice, various combinations of individual features are usually used to achieve the best possible results.

In the latter phase, various classification algorithms can be used, such as support vector machines (Shin, Chang, and Kim, 2010) or Gaussian mixture models (Ng et al., 2012; Ghaemmaghami, Dean, et al., 2015). In recent years, various deep neural networks architectures started to be employed more and more frequently, including fully connected feed-forward DNNs (Ryant, Liberman, and Yuan, 2013), convolutional neural networks (Saon et al., 2013) or recurrent neural networks (Hughes and Mierle, 2013; Eyben et al., 2013). More complex approaches, such as jointly trained DNNs (Q. Wang et al., 2015) or boosted DNNs (X. Zhang and D. Wang, 2014), have also been proposed. Moreover, in (Thomas et al., 2015) a combination of DNN and CNN is used. The output from a given classifier can also be smoothed to further improve the accuracy of the detection. Recently, various techniques, such as the Viterbi decoder (Ryant, Liberman, and Yuan, 2013; Gao et al., 2011) or weighted finite state transducers (Chung, S. J. Lee, and Y. Lee, 2013), have been applied for this purpose.

Most of the aforementioned works aim primarily at offline applications because applying speech activity detection in an online environment brings further restrictions on the system, such as low computational demands and latency. The ap-

proaches developed namely for the online task include, for example, conditional random fields (Gao et al., 2011) or accurate endpointing with expected pause duration (Liu, Hoffmeister, and Rastrow, 2015). Another approach in (Moattar and Homayounpour, 2009) utilizes short-term features.

There is a lack of standardized datasets. The most utilized one is QUT-NOISE-TIMIT (Dean et al., 2010) corpus. Unfortunately, a lot of papers opt for their own data making the comparison of different approaches much harder.

1.2 Speaker Change Point Detection

Similarly to speech activity detection, speaker change point detection (on extracted speech segments) is usually carried out in two consecutive stages: feature extraction and speaker segmentation itself.

In the first stage, a wide range of features has been utilized over the years. Simpler features, such as short-time energy (Meignier, Moraru, et al., 2006), zero-crossing rate (Lu and H. Zhang, 2002) or pitch (Lu and H. Zhang, 2005), were successfully employed. MFCCs (Sinha et al., 2005; Lu, H. Zhang, and Jiang, 2002) are/were probably the most used features even though they were designed for a different task. Line spectrum pairs (LSP) (Lu, H. Zhang, and Jiang, 2002; Lu and H. Zhang, 2005) and perceptual linear prediction cepstral coefficients (PLP) (Tranter et al., 2004; Chu, Tang, and Huang, 2009) were quite popular as well. Recently, a focus has shifted to crafting more complex features capturing more speaker specific information. Nowadays, i-vectors (Zhu and Pelecanos, 2016; Neri et al., 2017) are the go-to features for most of the state-of-the-art systems. DNNs have also been used to extract features (K. Chen and Salman, 2011; Yella and Stolcke, 2015). Novel features, d-vectors, were presented in (R. Wang et al., 2017) yielding very good results. The latest trend are speaker embeddings (C. Li et al., 2017; Bredin, 2017; Jati and Georgiou, 2017), which are primarily designed for end-to-end systems. In practice, the best results are often achieved by a mixture of aforementioned features.

The speaker segmentation approaches can be divided into three main categories: metric-based, model-based and hybrid-based. The metric-based approaches require a distance measure to be defined first. After that, two adjacent windows are shifted along the recording, and the distance between them is computed. If the distance is greater than a predefined threshold, a change point is labeled. The most commonly used distance measures are Euclidean distance (Lu and H. Zhang, 2002), Bayesian information criterion (S. S. Chen and Gopalakrishnan, 1998; Cettolo, Vescovi, and Rizzi, 2005), generalized likelihood ratio (Gish, Siu, and Rohlicek, 1991), Gaussian divergence (Barras et al., 2006) and Kullback-Leibler divergence (Siegler et al., 1997). Support vector machines (Fergani, Davy, and Houacine, 2008) were also successfully employed. On the plus side, these approaches do not require any prior knowledge about the recording (number of speakers, signal characteristics, etc.). Unfortunately, threshold tuning and frequent consecutive changes in a short span of time are the main obstacles of these methods. The model-based approaches utilize different models trained from labeled audio data (prior knowledge).

These models are then employed to detect speaker change points when the change from one speaker to another happens. In literature, the common approaches are HMMs (Meignier, Bonastre, and Igounet, 2001) and GMMs (Magrin-Chagnolleau, Rosenberg, and Parthasarathy, 1999; Malegaonkar, Ariyaeeinia, and Sivakumaran, 2007). Eigenvoice-based models (Castaldo et al., 2008; Desplanques, Demuynck, and Martens, 2015) can outperform GMM-based models. With the recent advances in deep learning in mind, systems based on DNNs (Gupta, 2015), CNNs (Hrúz and Zajíc, 2017) and bidirectional long short-term memory RNNs (Yin, Bredin, and Barras, 2017) yield state-of-the-art results. Hybrid-based segmentation approaches combine the metric and model-based approaches to exploit the advantages from both worlds (e.g. (Moraru et al., 2004)).

The restrictions of applying speaker change point detection in online environment (e.g., low computational demands and low latency) result in significantly lower number of published work. Furthermore, most of this work has been done for the task of speaker diarization. The common approaches are based on GMMs (Markov and Nakamura, 2007; Geiger, Wallhoff, and Rigoll, 2010; Soldi, Beaugeant, and Evans, 2015). In (Dimitriadis and Fousek, 2017) the authors explored Bayesian information criterion, i-vector features and within class covariance normalization for online application. The use of i-vectors was also investigated in (Zhu and Pelecanos, 2016).

In literature, there are several commonly used datasets for training and evaluation of speaker segmentation. One of the first often employed dataset is Hub-4 (Stern, 1997). The French datasets ESTER (Galliano et al., 2006), ETAPE (Galibert, Leixa, et al., 2014) and REPERE (Galibert and Kahn, 2013) are also commonly utilized. Speaker segmentation can also be evaluated on multilingual database COST278 (Zibert et al., 2005). Recently, The Speakers in the Wild dataset (McLaren et al., 2016) has been published. Some other worth mentioning datasets are CALLHOME and NIST SRE. However, most of the published works report the results on only one selected dataset making the comparisons harder.

1.3 Existing Systems

ALIZE Speaker Recognition toolkit (Bonastre, Wils, and Meignier, 2005; Larcher et al., 2013) is an open-source tool primarily designed for speaker recognition. As such, it also provides support for speaker segmentation (based on HMMs) (Bozonnet, Evans, and Fredouille, 2010). LIUM Speaker Diarization (Meignier and Merlin, 2010; Rouvier et al., 2013) is probably the most known toolkit. It was originally developed for French ESTER2 evaluation campaign for diarization of broadcast news, and it provides tools for feature extraction (MFCCs), speech activity detection (HMMs), gender detection, speaker segmentation (GMMs, BIC) and speaker clustering. It also comes with pretrained models for immediate use. DiarTK (Vijayasenan and Valente, 2012) is another toolkit based on GMMs focused on multistream speaker diarization. Pyannote is a brand new option providing scripts for speech activity detection (Yin, Bredin, and Barras, 2017), speaker change point de-

tection (Yin, Bredin, and Barras, 2017) and speaker embeddings (with pretrained models) (Bredin, 2017). It is based on long short-term memory recurrent neural networks, and it yields very promising results. Newly, speaker segmentation based on deep neural networks is being worked on to be added to Kaldi toolkit (Povey et al., 2011).

Other notable systems, such as CMU Segmentation toolkit, AudioSeq or SHoUT toolkit, can be utilized as well, but their performance is usually outperformed by new counterparts.

2 Motivation and Goals

The recent advances in deep learning resulted in a novelty approach to speech recognition (Dahl et al., 2012), which yielded state-of-the-art results by a large margin over the previously conventional approaches. This success provoked further research of deep neural networks and their application to a different range of application. In case of this thesis, it is for the task of speaker change point detection.

SpeechLab (Laboratory of Computer Speech Processing) of Technical University of Liberec (TUL) has been focusing on speech processing and speech recognition for a long time. The TVR monitoring system developed here at SpeechLab@TUL carries out 24/7 transcription of radio and TV broadcasts in a range of different languages. In peak hours (during the day), it transcribes up to 120 streams in parallel in real-time. During the non-prime hours (mostly in night), it is still processing at least 20 online streams every second. The day average ranges from 60 to 80 simultaneously transcribed online streams. Approximately 133 days (3,196 hours or 75GB) of recordings are being processed every day. The biggest chunk of every day transcribed streams form Polish (80 broadcasts monitored), Czech (47) and Slovak (12) broadcasts. However, a wider range of Slavic languages, such as Russian (approximately 20 broadcast monitored), Bulgarian (20), Croatian (10) or Serbian (10) and more, are being transcribed as well.

An integration of speaker change point detection approach (operating in two consecutive phases: speech activity detection and speaker segmentation) to this existing system would be beneficial to SpeechLab@TUL for many reasons. Firstly, speech activity detection would be used as a preprocessor for online streams to run the transcriber only on speech segments. This should result in significant reduction in processing time, and it should ease the CPU load as well (if the stream contains a lot of non-speech segments (e.g. music radios)). It should also yield a better accuracy of transcriptions as the non-speech parts are omitted from being transcribed. Secondly, the detected speech segments would be used as inputs to speaker segmentation to find transitions between two speakers. The created labels would ease the handling of online streams as it would provide additional information, and it would segment the streams to smaller speaker homogeneous chunks, which could be easily further utilized. These chunks form a good starting point for a full diarization system, which could be then extended to speaker verification & identification systems to provide the transcribed recordings with even more valuable information. The detected segments could also be used as training data for future speaker adaptive approaches to speech recognition. Lastly, SpeechLab@TUL is always trying to improve the TVR monitoring system with additional functionality, especially based on state-of-the-art-technologies, and is planning to do so in future as well. Speaker segmentation is considered as one of the wanted additional functionality.

An integration of any of the existing systems (see Sect. 1.3) would be a tough task as they are not a very good fit for the requirements of SpeechLab@TUL. They are quite often fine-tuned to very specific conditions (telephone conversations, broad-

casts (French - LIUM Speaker Diarization), etc.), which may or may not be suitable for TVR monitoring system. However, it would most likely result in a need of training new models on proper data. Nextly, except for Pyannote, which was released recently in 2017, all of the systems are built on overcome technologies (mostly GMM-based), and they do not yield state-of-the-art results. An approach based on deep learning would fit philosophy of SpeechLab@TUL more as the transcriber is based on deep neural networks as well. And most importantly, none of the systems is primarily designed for online use. The required input is an audio file or a parametrized signal, and some of the systems even require multiple passes through data (e.g. LIUM Speaker Diarization). A lot of additional work would be needed to get the systems functioning for online streams. It might not even be possible for some systems. Lastly, the TVR monitoring system is a distributed application. As such, it is distributed in a docker image. This brings further requirements, such as good scalability or fast and stable implementation. These facts result in a need of speech change point detection approach developed specifically for the TVR monitoring system.

The main goal of this work is thus to develop a speaker change point detection approach that:

1. utilizes state-of-the-art techniques including namely DNNs,
2. yields at least comparative results with already existing methods,
3. operates in two or more consecutive phases, where the first one allows for robust speech/non-speech detection,
4. operates in an online mode with a low latency in order to process real-time streams,
5. can be integrated into an existing TVR monitoring system developed at SpeechLab@TUL.

3 Current Status

The final thesis is still a work in progress. The progress can be divided into two successive tasks as the proposed speaker change point approach operates in two consecutive phases as well:

- Speech Activity Detection - **completed and published**
- Speaker Change Point Detection - **in progress**

3.1 Speech Activity Detection

The proposed speech activity detection approach was designed in several consecutive steps. After presenting the employed evaluation metrics and development data, this section follows these steps from the initial design to the final proposed approach (see Sect. 3.1.6). Lastly, the performance of the final proposed SAD approach was evaluated on standardized QUT-NOISE-TIMIT dataset as well as in real speech transcription system.

The proposed speech activity detection approach was presented at SIGMAP 2016 conference (Mateju, Cerva, and Zdansky, 2016), ICASSP 2017 conference (Mateju, Cerva, Zdansky, and Malek, 2017) and in E-Business and Telecommunications (Mateju, Cerva, and Zdansky, 2017).

3.1.1 Metrics Used for Evaluation

In total, seven metrics were employed for evaluation of proposed speech activity detection approach. These metrics can be divided into three main groups each focusing on different aspect of the task: overall accuracy metrics, change point quality measures and performance measures.

Overall Accuracy Metrics

Overall accuracy metrics focus on a precision of speech activity detection on a frame level. Four different metrics are within this group.

Frame error rate, the first metric, is defined as follows:

$$FER[\%] = \frac{M}{N} * 100 \ , \tag{3.1}$$

where M is the number of non-matching frames between reference and decoded output, and N is the total number of frames in the reference.

The following two metrics, miss rate and false alarm rate (Ryant, Liberman, and Yuan, 2013), represent relevance measures, specifically false negatives and false positives. The rest of the relevance measures is not reported in this work as they are complementary to the presented ones.

Miss rate (false negatives) is defined as follows:

$$MR[\%] = \frac{M_{\text{speech}}}{N_{\text{speech}}} * 100 \ , \tag{3.2}$$

where M_{speech} is the number of misclassified speech frames, and N_{speech} is the total number of reference speech frames.

False alarm rate (false positives) can be expressed as:

$$FAR[\%] = \frac{M_{\text{non-speech}}}{N_{\text{non-speech}}} * 100 \ , \tag{3.3}$$

where $M_{\text{non-speech}}$ is the number of misclassified non-speech frames, and $N_{\text{non-speech}}$ is the total number of reference non-speech frames.

The last metric, half-total error rate, is defined as equal-weighted average of MR and FAR:

$$HTER[\%] = \frac{MR + FAR}{2} \ . \tag{3.4}$$

This metric is only evaluated for experiments on QUT-NOISE-TIMIT dataset (see Sect. 3.1.7) to compare the achieved results with literature.

Note that the optimal speech activity detection approach should minimize the miss rate while keeping the false alarm rate fairly low. The reason is that the desired speaker segmentation/transcription system should get all speech frames with only limited amount of non-speech events added.

Change Point Quality Measures

Change point quality measures, as the name suggests, focus on a precision of detected change points between speech/non-speech segments. For this task, two metrics, F-value and $\delta_{2/3}$, were employed.

To evaluate the quality of the change point detection, the detected (computed) boundaries and the reference boundaries have to be aligned at first (Räsänen, Laine, and Altosaar, 2009). After that, hits (H), insertions (I) and deletions (D) can be defined. A change point is considered as a hit if the detected and reference boundaries are nearest to each other (timewise) and within certain time threshold. Insertions and deletions then mark errors. If a detected boundary do not match any of the reference boundaries, it is tagged as insertion. Similarly, if a reference boundary is not matched by any of the detected boundaries, it is marked as deletion. The value of the threshold hugely depends on the task itself and in practise differs widely. Within the scope of this work, it was set to 1 second.

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 boundaries and the number of detected boundaries:

$$P[\%] = \frac{H}{H + I} * 100 \ , \tag{3.5}$$

while recall is expressed as a ratio between the number of correctly detected boundaries and the number of boundaries in reference:

$$R[\%] = \frac{H}{H+D} * 100 \ . \tag{3.6}$$

Finally, F-value, the metric reported in this work, can be computed as:

$$F{-}value[\%] = \frac{2 * R * P}{R+P} \ . \tag{3.7}$$

Given the correctly detected boundaries (hits), it is also possible to calculate an error value for each hit (in seconds) and sort all the hits according to these calculated values in ascending order. In this work, the measure $\delta_{2/3}$ is utilized, which expresses (in seconds) the maximal error of the alignment for first two-thirds of the sorted (best) hits. Note that $\delta_{2/3}$ should be as low as possible to provide the speaker segmentation with precisely determined speech segments to process.

Performance Measures

Two metrics, latency and real-time factor, were evaluated to monitor the performance of proposed speech activity detection approach in online environment.

Latency is defined as an average time between the detected change point and the moment the decoder outputs the change point label. Keeping this value as low as possible is necessary in online applications.

The second metric, real-time factor, measures the speed of decoding:

$$RTF = \frac{T}{PT} \ , \tag{3.8}$$

where PT is the processing time of decoding, and T is the duration of the recording. Lowering RTF results in speeding up the decoding.

3.1.2 Data Used for Development

The data used for development and evaluation consisted of 6 hours of TV and radio recordings in several Slavic languages (Czech, Slovak, Polish and Russian). It contained not only clean speech segments but also segments with music, background noises, jingles and/or advertisements. Annotations of this data were obtained in a two step process. At first, speech and non-speech labels were produced automatically by the baseline DNN-based SAD approach (see Sect. 3.1.3). These obtained labels were then corrected and fine-tuned by hand. In total, 70% of all frames were marked as speech ones. Note that the performance on clean speech and non-speech (music) data was reported in (Mateju, Cerva, and Zdansky, 2016).

3.1.3 Baseline DNN-Based Approach

The baseline approach employed a feed-forward deep neural network with a binary output (speech/non-speech) for each frame (i.e., without any smoothing). In total,

67 hours of recordings were utilized for DNN training. The speech class was 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 system developed at SpeechLab@TUL. The non-speech class was modelled by 30 hours of music of different genres with addition of 7 hours of non-speech events/noises.

The initial DNN hyper-parameters were set to:

- 5 hidden layers;
- 128 neurons per hidden layer;
- ReLU activation function;
- mini-batches size of 1024;
- 0.08 learning rate;
- 10 epochs.

The features extracted from training data were:

- 39-dimensional log filter banks;
- concatenation of 25 previous frames, the current frame and 25 following frames;
- local normalized within one second window.

The tuning of hyper-parameters and features was discussed in (Mateju, Cerva, and Zdansky, 2017).

The accuracy of the baseline approach is summarized in first row of Table 3.1 (Baseline DNN-based). It is evident that it missed approximately 4% of speech segments. This fact would affect the accuracy of the speech transcription system negatively as the segments incorrectly marked as non-speech would not get transcribed. Another problem of the baseline detector is the time precision of the change-point detection: the achieved value of $\delta_{2/3}$ is 0.42 s. 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 speech/non-speech segments with a very short duration of one or two frames. In reality, every speech/non-speech segment usually lasts for at least several frames.

Note that all deep neural networks (i.e., for all experiments) were trained on GPU using the torch framework[1]. The training scripts are available at author's GitHub[2] for everyone to use/download.

[1] http://torch.ch
[2] https://github.com/1shark1/nnet

Table 3.1: Summarized results of the proposed SAD approach described in Sect. 3.1.

Approach	FER [%]	MR [%]	FAR [%]	F-value [%]	$\delta_{2/3}$ [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 smooth.	2.4	0.5	7.2	52.7	0.26

3.1.4 Smoothing the Output from DNN

As stated in the previous section, the binary outputs of DNN had to be smoothed out to suppress the frequent changes between extremely short speech/non-speech segments. For this task, weighted finite state transducers (and consecutively OpenFst library[3]) were utilized.

The final decoding scheme is composed of two transducers. The first one, as depicted in Fig. 3.1, models the input signal. The second one, the transduction model, represents the smoothing algorithm (see Fig. 3.2). It consists of three states. The first state, denoted by 0, is the initial state. The speech/non-speech labels are emitted by transitions between states 1 and 2. These transitions are penalized by penalty factors P1 and P2. Their values (500 and 500) were experimentally tuned on different dataset.

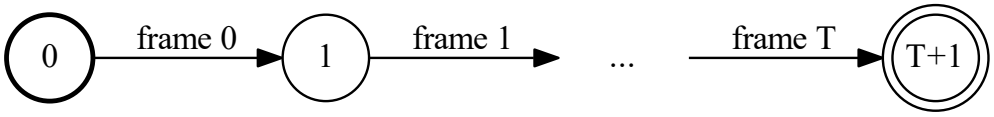

Figure 3.1: Weighted finite state transducer representing the input signal.

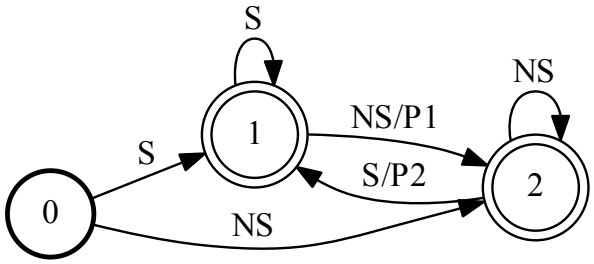

Figure 3.2: Weighted finite state transducer representing the basic smoothing model without any context for SAD.

Given the two above described transducers, the decoding process is performed using on-the-fly composition of the transduction and the input model of an unknown size. This is feasible since the input is considered to be a linear-topology,

[3]http://www.openfst.org/twiki/bin/view/FST/WebHome

un-weighted, epsilon-free acceptor. After each composition step, the shortest-path (considering tropical semiring) 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 certain, it is denoted as a temporary output (i.e., it can still be altered later in the process).

The influence of smoothing the outputs of DNN on results can be observed in second row of Table 3.1 (+ Basic smoothing). The results show an overall significant boost in all metrics. For example, F-value 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 s to 0.27 s.

3.1.5 Using Artificial Training Data

The level of MR yielded so far (around 2%) would still lead to a small increase in word error rate of the transcription system (e.g. from 13% to 14%) as the misclassified speech frames would be omitted from transcription. Upon closer inspection, most of the misclassified speech frames are segments with background noise. This was caused by utilizing only speech data recorded in clean conditions (i.e., without any background noise) during DNN training.

To resolve this issue, training data containing various non-speech events in the background (e.g., music, jingles) were required. 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. 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 SNR (which was also selected randomly from an interval between -30 dB and 50 dB).

The labels of this artificial data were created automatically. If SNR of a recording was higher than a defined threshold (0 dB), the recording was labeled as speech. In the opposite case, the recording was annotated as non-speech.

The results are summarized in third row of Table 3.1 (+ Artificial training data). It is evident that utilizing this artificial data 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 s to 0.34 s. Due to these issues, a further refinement of the smoothing algorithm was investigated.

3.1.6 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 original smoothing model (i.e., 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 (*end_S*) to the start state of non-speech (*start_NS*), and b) from the end state of non-speech (*end_NS*) to the start state of speech (*start_S*).

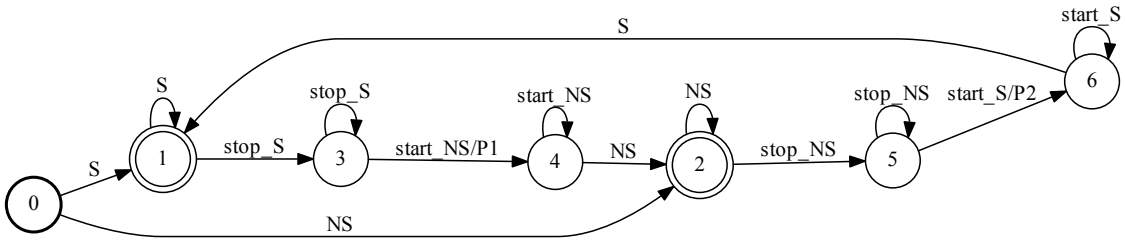

Figure 3.3: Weighted finite state transducer representing the context-based smoothing model for SAD.

The training data from Sect. 3.1.5 had to be modified to include transition between speech/non-speech events. At first, two recordings were chosen randomly, one speech and one non-speech. After that, these two recordings were concatenated in a random order. The final recording then contained one of the two possible transitions (i.e., from speech to non-speech or from non-speech to speech) and was labeled 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.

This approach, as the results show (see fourth row (Modified artificial data + context-based smoothing) in Table 3.1), a) addresses the issue of an increase in $\delta_{2/3}$, which has emerged due to the use of the artificial training data, b) significantly decreases FAR and c) increases F-measure. The value of $\delta_{2/3}$ was reduced from 0.34 s to 0.27 s. The negligible downside is a slight increase in miss rate (by 0.2%).

After achieving these satisfactory results, this speech activity detection approach was evaluated on standardized dataset.

3.1.7 Evaluation on QUT-NOISE-TIMIT Corpus

To compare the performance of the final proposed SAD approach (see Sect 3.1.6) with different approaches presented in literature, a standardized QUT-NOISE-TIMIT (Dean et al., 2010) corpus was utilized. The comparison was done with five approaches already presented in (Dean et al., 2010) and two novel approaches with even better, state-of-the-art, results (Wisdom et al., 2015; Ghaemmaghami, Dean, et al., 2015).

The five original SAD systems were: standardized VAD system ITU-T G.729 Annex B (Benyassine et al., 1997), standardized advanced front-end ETSI (J. Li et al., 2004), Sohn's likelihood ratio test (Sohn, Kim, and Sung, 1999), Ramirez's long-term spectral divergence (LTSD) (Ramírez et al., 2004) and GMM based approach with use of MFCC features (Dean et al., 2010). The newer approaches were voice activity detection using subband noncircularity (SNC) (Wisdom et al., 2015) and complete-linkage clustering (CLC) for VAD (Ghaemmaghami, Dean, et al., 2015).

QUT-NOISE-TIMIT Corpus

As the authors stated, the main idea behind the creation of QUT-NOISE-TIMIT corpus (Dean et al., 2010) was a need for a standardized corpus for training and testing of speech activity detection in various target environments and in different SNR conditions. For this purpose, they gathered more than 10 hours of background noise across 10 different unique locations to a corpus called QUT-NOISE. These background noises covered five different but common scenarios (specifically cafe, home, street, car and reverb). Each scenario was also composed of two different source locations:

- cafe - outdoor cafe environment or indoor shopping food-court;
- home - kitchen or living room;
- street - inner-city or outer-city traffic-light controlled intersections;
- car - windows down or up;
- reverb - indoor pool or partially enclosed carpark.

The QUT-NOISE background noises were mixed with a clean speech from TIMIT corpus (Fisher, Doddington, and Goudie-Marshall, 1986) creating 600 hours of new recordings with various amount of speech segments, length (60 or 120 seconds) and SNR level (-10, -5, 0, 5, 10 or 15 dB). These new recordings then formed the standardized QUT-NOISE-TIMIT corpus. After that, the final corpus was evenly split into two groups (A and B) to provide training and testing subsets.

Evaluation Protocol

The authors of QUT-NOISE-TIMIT also provided an evaluation protocol, which was than followed by others as well (Wisdom et al., 2015; Ghaemmaghami, Dean, et al., 2015). During the training phase, no information of target scenario was

Table 3.2: Performance of the proposed SAD approach on QUT-NOISE-TIMIT.

Conditions	HTER [%]	FER [%]	MR [%]	FAR [%]	F-value [%]	$\delta_{2/3}$ [s]
Low noise	2.7	2.6	2.2	3.0	85.1	0.05
Medium n.	5.8	5.8	5.8	5.8	65.2	0.11
High n.	17.0	16.4	24.0	10.0	33.1	0.22

given to the system. The only available prior knowledge was the SNR level of target environment: low noise (10, 15 dB), medium noise (0, 5 dB) or high noise (-10, -5 dB). For each of these target environments, group A was used for training and group B for testing and vice-versa. The decoded segments were then compared with QUT-NOISE-TIMIT ground truth labels, and miss rate, false alarm rate and half-total error rate were evaluated (see Sect. 3.1.1).

The proposed SAD approach followed this evaluation protocol, and it was trained as described in Sect. 3.1.6, of course, with the exception of not using the artificial training data (i.e., only the data from QUT-NOISE-TIMIT were utilized).

Low Noise Conditions

The experiment in low noise conditions was based on recordings with SNR level of 10 and 15 dB. The results are depicted in Fig. 3.4. As the results show, the proposed SAD approach outperformed all other systems by a fair margin. The absolute reduction in HTER was more than 2% over the second best complete-linkage clustering approach. The achieved value of HTER was 2.6%, the other metrics are given in first row of Table 3.2. Note that the precision of the detected change points was also very good (F-value 85.1% and $\delta_{2/3}$ 0.05 s). The proposed SAD approach thus achieved state-of-the-art results in low noise conditions.

Medium Noise Conditions

Recordings with SNR level of 0 and 5 dB were utilized for experiment in medium noise conditions. Figure 3.5 depicts the results. Similarly to experiment conducted in low noise conditions, the proposed SAD approach yielded the best results, outperformed the other systems and thus reached the state-of-the-art results. Once again, the absolute reduction in HTER was over 2% (the achieved value was 5.8%, see second row of Table 3.2) over the second best CLC system. The worsen conditions caused an increase of over 3% in HTER for the proposed SAD approach.

Hard Noise Conditions

The hardest conditions to segment were based on recordings with SNR level of -10 and -5 dB. The comparison of performance of various SAD systems is depicted in Fig. 3.6. The proposed SAD approach was outperformed by approximately 2% in HTER by complete linkage clustering. However, it still outperformed the rest of the systems by a fair margin. The achieved HTER was 17% (an increase of over 11% over the medium noise conditions). The rest of the metrics are shown in third row of Table 3.2. Unfortunately, most of its errors were caused by omitted speech (i.e.,

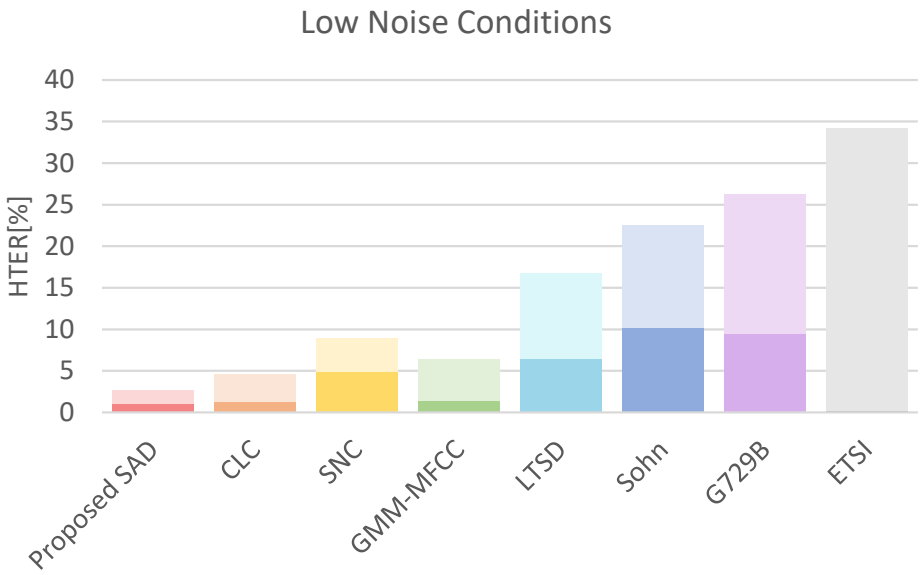

Figure 3.4: Comparison of results of various SAD approaches in low noise conditions on QUT-NOISE-TIMIT dataset. The contribution of MR and FAR to HTER bars is displayed by darker and lighter shades, respectively.

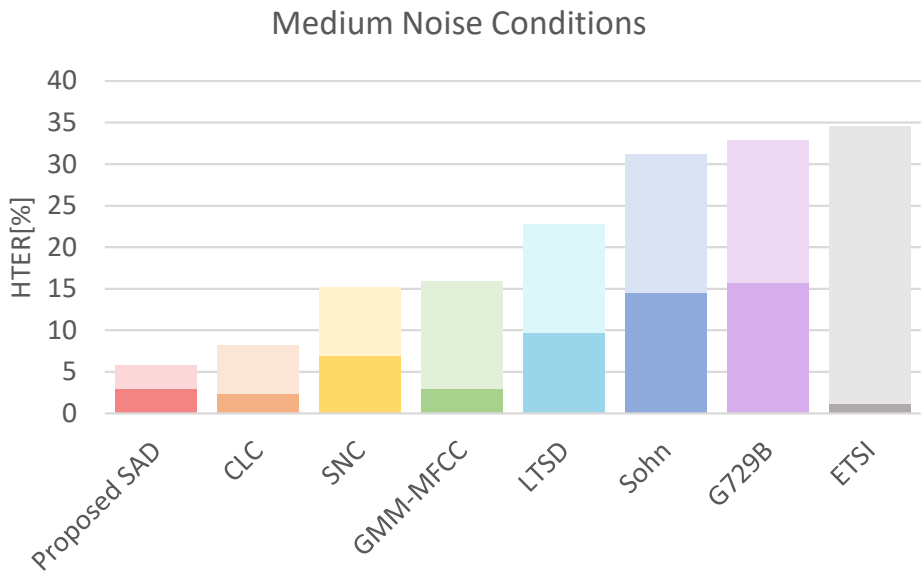

Figure 3.5: Comparison of results of various SAD approaches in medium noise conditions on QUT-NOISE-TIMIT dataset. The contribution of MR and FAR to HTER bars is displayed by darker and lighter shades, respectively.

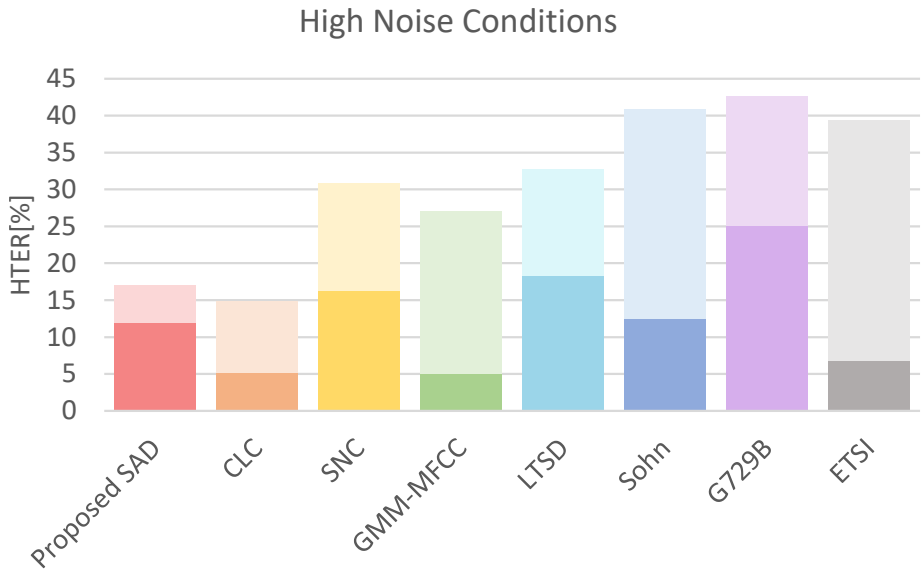

Figure 3.6: Comparison of results of various SAD approaches in high noise conditions on QUT-NOISE-TIMIT dataset. The contribution of MR and FAR to HTER bars is displayed by darker and lighter shades, respectively.

higher miss rate). However, the proposed SAD approach was not designed and fine-tuned for such low SNR conditions.

3.1.8 Performance Details

A performance of the proposed speech activity detection approach was closely monitored so that it could be integrated into a real speech transcription system without any issues. The proposed SAD approach averaged real-time factor of 0.01 and latency around 2 seconds (these values were measured using processor Intel Core i7-3770K @ 3.50GHz). This performance allows for its seamless use in real-time processing without any major delay.

3.1.9 Evaluation of Proposed SAD Approach in a Real Speech Transcription System

Given all the findings, results from previous experiments and fast performance, the final proposed speech activity detection approach (i.e., with context-based smoothing model) was integrated into TVR monitoring system developed at SpeechLab@TUL and thus evaluated in a real speech transcription system.

For this purpose, two test sets of Czech broadcasts were utilized (see Table 3.3). The first set, recordings from local TV news channel, contained 22204 words, and approximately 60% of its 4 hour length was labeled as speech. The other set represented a local music radio station. Its length was 8 hours (7212 words), and only 10% of its content was considered as speech.

The transcription system developed at SpeechLab@TUL employed an acoustic

Table 3.3: Information about test sets for speech transcription.

Test set	Duration	Speech [%]	Words
Live news TV channel	4 hours	60%	22204
Local radio station	8 hours	10%	7212

model based on a Hidden Markov Model - Deep Neural Network (HMM-DNN) hybrid architecture (Dahl et al., 2012), where the baseline Gaussian Mixture Model (GMM) was trained as context-dependent, speaker-independent and contained 3886 physical states. The acoustic model was trained on phonetically annotated 270 hours of clean speech recordings of Czech. The hyper-parameters of DNN could be summed to:

- 5 hidden layers;
- decreasing number of neurons per layer (1024-1024-768-768-512);
- ReLU activation function;
- mini-batches size of 1024;
- learning rate 0.08;
- 35 epochs.

The features were:

- 39-dimensional log filter banks;
- concatenation of 5 previous frames, the current frame and 5 following frames;
- local normalized within one-second window.

Note that fine-tuning of DNN hyper-parameters (used within this experiment) was published in (Mateju, Cerva, and Zdansky, 2015) and later on further extended.

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 N-grams. For practical reasons (mainly with respect to the very large vocabulary size), the system used bigrams. However, 20 percent of all "word-pairs" actually included sequences containing three or more words, as the lexicon contains 4,000 multi-word collocations. The unseen bigrams were backed-off by Kneser-Ney smoothing (Kneser and Ney, 1995).

Experimental Results

Within the scope of experimental evaluation, both test sets were transcribed a) with and b) without the use of the final speech activity detection approach. The results are presented in Table 3.4, which contains values of word error rate and correctness (i.e., presents the accuracy of transcription). To measure computational demands with and without applying SAD, values of real-time factor are also presented.

Table 3.4: Evaluation of the proposed SAD approach in a real speech transcription system.

Test set	live news TV channel		local radio station	
SAD module	Yes	No	Yes	No
WER [%]	12.4	12.7	14.0	17.9
correctness [%]	89.7	89.7	88.5	88.4
RTF	0.42	0.77	0.08	0.83

The results show that the utilization of the proposed speech activity detection approach was beneficial on both test sets. The yielded improvements in WER and correctness proved that the proposed SAD approach omitted hardly any speech parts and even decreased the number of insertions (from non-speech parts). The real-time factor was almost two times and more than ten times lower for local TV and radio broadcast test sets, respectively. The decrease was, of course, more significant for the radio set because it contained significantly more non-speech events (e.g. music). This resulted in huge savings in processing time and computational resources.

A small reminder, RTF of the proposed SAD approach is 0.01 and its latency is around 2 seconds (see Sect. 3.1.8). Given these numbers and the fact that RTF of the transcription system is around 1, it's perfectly feasible to use the proposed speech activity detection approach in TVR monitoring system developed at SpeechLab@TUL without any major delay.

Thanks to all these findings, the proposed speech activity detection approach is also perfectly capable of working as the first phase of speaker change point detection to identify speech segments in recordings.

3.2 Speaker Change Point Detection

The proposed speaker change point detection approach is also being designed in several consecutive steps. These steps, from initial approach to the current one, are heavily discussed within this section. The evaluation metrics and development data are also presented.

Note that the speech segments obtained by the proposed speech activity detection approach are utilized as inputs to speaker segmentation.

Also note that the speaker change point detection approach is still in development, and the first publication is planned for next year.

3.2.1 Metrics Used for Evaluation

In total, 4 metrics (already presented for speech activity detection), were utilized to evaluate the performance of speaker change point detection. Specifically, the metrics are precision, recall and derived from them, F-value. Finally, the last metric is $\delta_{2/3}$. The definitions and further information can be seen in Sect. 3.1.1 in Change Point Quality Measures paragraph.

3.2.2 Data Used for Development

A small subset of standardized dataset COST278 (Zibert et al., 2005), specifically Czech test data, was used for evaluation and development of speaker segmentation. This test set consisted of four recordings of different Czech broadcasts (ČT1, Nova & Prima) in a total length of hour and half. It not only contained clean speech segments but also segments with background noise and jingles. The annotations provided with COST278 dataset were utilized (these annotations can even be further used, e.g., for speaker verification & identification tasks). In total, there were 399 labeled transitions from one speaker to another.

3.2.3 Baseline DNN-Based Approach with Smoothing

The baseline approach utilized a deep neural network as a binary classifier (change point/no change point) and weighted finite state transducers as a decoder to smooth the outputs of DNN. The smoothing was integrated into the baseline approach straight away as it was necessary for proposed speech activity detection approach to yield at least decent results (as proved in Sect. 3.1.4).

A large dataset of manually annotated (or automatically annotated and corrected by hand) broadcasts of Czech TV/radio shows from 2009 to 2014 was utilized to extract training data. In these broadcasts, transitions between two speakers were found and extracted (only if the duration of silence during the speaker transition was shorter than one second (longer silence should get detected by SAD)). In total, 30 hours of recordings (with the average length of one recording being around 5 seconds) were gathered. This resulted in 20,000 speaker transitions, which could be divided into 4 groups (the transition from female to female, female to male, male to female and male to male) each represented by 5,000 change points.

The annotations for this extracted data were generated in an automated way. The frame corresponding to the actual change point as well as safety collar frames around it were labeled as change point. This safety collar was set to 1 second (100 frames). This is due to the fact that determining the precise change point is quite often an ambiguous task (silence, crosstalk, etc.), and to provide the network with more information about the transition from one speaker to another. The rest of the recording was labeled as segments without transition. The annotation of one recording is depicted in Fig. 3.7.

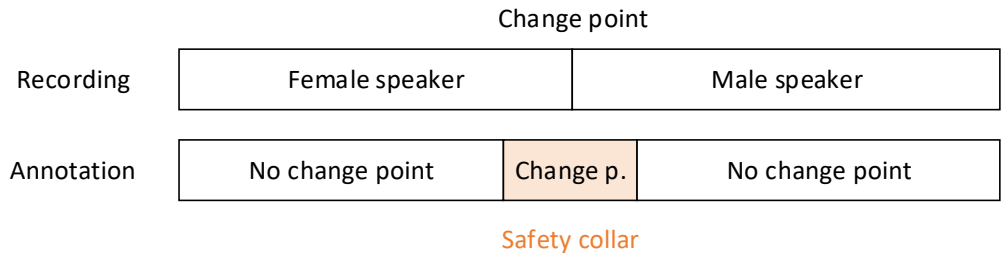

Figure 3.7: Example of annotation of recording from training data for SCP.

The hyper-parameters of deep neural network utilizing this data were set to:

- 2 hidden layers;
- 64 neurons per hidden layer;
- ReLU activation function;
- mini-batches size of 1024;
- 0.08 learning rate;
- 15 epochs.

The extracted features were:

- 39-dimensional log filter banks;
- concatenation of 100 previous frames, the current frame and 100 following frames;
- no local normalized.

The decoding was done as described in Sect 3.1.4 with the exception of applying different transduction model (slightly modified to be more suitable for speaker change point detection). This transduction model (depicted in Fig. 3.8) consists of two states (denoted 0 and 1). The transitions between states 0/1 emit labels start/end change point with a time mark. The final change points are then in the middle of start and end change point time marks. The transitions are also weighted by penalty factors P1 and P2. Note that their values (25 and 25) were tuned on different dataset.

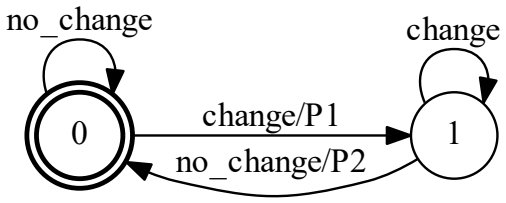

Figure 3.8: Weighted finite state transducer representing the basic smoothing model for SCP.

The results of the baseline approach are summarized in first row of Table 3.5. The achieved results provide a decent starting point, but especially the precision is not particularly good. This resulted in a need of different transduction model, which would be more suitable for the task of speaker segmentation.

3.2.4 Improved Smoothing with Forced Duration

The improved smoothing scheme was designed to reflect the annotation style of training data and fit the task of speaker change point detection more. As described in previous section, 1 second long window (100 frames) around the actual change

Table 3.5: Summarized results of the proposed SCP approach described in Sect. 3.2.

Approach	Precision [%]	Recall [%]	F-value [%]	$\delta_{2/3}$ [s]
Baseline DNN-based + Basic smoothing	46.9	72.8	57.1	0.23
+ Forced duration smoothing	48.7	76.0	59.3	0.22
+ Artificial data	65.4	73.9	69.4	0.20

point was labeled as change point during the training. However, during the decoding, the duration of the transition between speakers was decided by the decoder itself and varied greatly. Within this experiment, the duration of this transition was forced to be precisely 1 second.

To reflect this concept, the transduction model had to be modified. The new scheme is depicted in Fig. 3.9. The model consists of two main states (0 and 1) and 98 forced transition states (depicted as …; the number of the states corresponds to the forced duration of the transition). When a transition from one speaker to another happens, the decoder has to go from state 0 through transition states to state 1; the change point label is output here; and back through the rest of the transition states to state 0, where it idles until next transition happens. The transitions are as usual penalized by factors P1 and P2 (weights being set to 25 and 25).

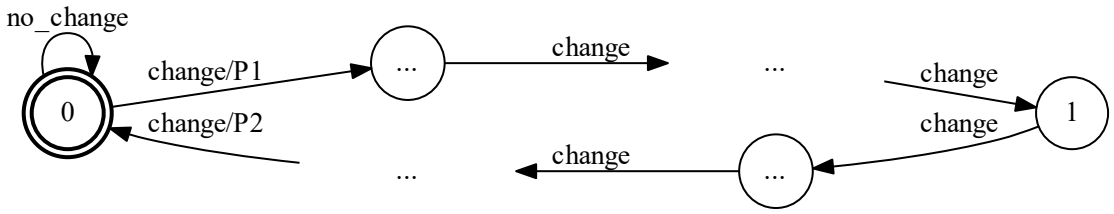

Figure 3.9: Weighted finite state transducer representing the smoothing model with forced duration for SCP.

The results are presented in second row of Table 3.5 (+ Forced duration smoothing). An overall slight improvement in all observed metrics was achieved (e.g. F-value increased from 57.1% to 59.3%). However, the results could still be hugely improved.

3.2.5 Adding Artificial Data

After evaluating the performance of the proposed approach on selected recordings (from different dataset), two types of errors were more prominent. The first common error was caused by quick transitions between speakers (without any silence) usually caused by artificial cuts (e.g. in news). These change points got consistently ignored by the decoder. The other common issue were longer silences (more than 0.5 s) in a speaker homogeneous segments. These longer silences were usually caused by a deep breath or hesitation, and they forced the decoder to output a false speaker change point. This issue is not that problematic as it could be fixed later on with speaker

clustering. However, both of these errors boil down to the fact that the deep neural network did not see such data during training.

For this reason, the training data presented in Sect. 3.2.3 was enriched by 20 additional hours. To deal with the first issue, approximately 10 hours of recordings containing quick cuts between speakers were prepared by artificially joining recordings of different speakers in quick succession. The labels were generated in the same way as for the original data. In total, 14,340 transitions were added to DNN training with uniform distribution between transition types (female-female, female-male, male-female, male-male). To solve the second issue, 10 hours of speaker homogeneous segments (including longer silences) were added to training. This whole 10 hours were labeled as segments with no change point. Within this experiment, DNN and WFST remained the same as presented in previous sections.

By utilizing this additional data, the results improved significantly (see third row of Table 3.5). The precision of change point detection increased from 48.7% to 65.4%, F-value got boosted from 59.3% to 69.4%, and even the speaker change point placement improved ($\delta_{2/3}$ was reduced from 0.22 s to 0.2 s). The only downside is a small decrease in recall (approximately 2%). However, this issue is easily overweighted by the overall improvements.

Even though the results improved significantly, there is still a room for further improvements before integrating the speaker change point detection into TVR monitoring system developed at SpeechLab@TUL. The design of the SCP approach is still being researched. As of right now, the results are fairly comparable with LIUM Speaker Diarization toolkit (without clustering).

4 Conclusions

This thesis statement presents the current progress of final thesis focused on a task of speaker change point detection. The main goal is/was to design a novel approach based on state-of-the-art technologies (namely DNNs), which can be integrated into existing TVR monitoring system developed at SpeechLab@TUL. For this reason, it has to support an online mode operating with low latency to process real-time streams. The final speaker segmentation operates in two consecutive phases:

- Speech Activity Detection - **completed and published**

  The proposed speech activity detection approach is based on feed-forward deep neural network and weighted finite state transducers. The DNN is used as a speech/non-speech classifier, while the context-based WFST decoder smooths the outputs. The network is trained on data artificially created by mixing speech and non-speech recordings at various levels of SNR. This design yields state-of-the-art results in low and medium noise conditions on standardized QUT-NOISE-TIMIT dataset. It is also suitable for online use as it operates with low real-time factor as well as low latency.

  The proposed speech activity detection approach is now fully integrated into the TVR monitoring system developed at SpeechLab@TUL. Last month, approximately 4,130 days (99,100 hours or 2.3TB) of recordings were transcribed in processing time of 1,333 days (32,000 hours). Considering the real-time factor of the speech transcriber is around 1, the utilization of SAD as preprocessor resulted in a significant saving of processing time. Approximately two thirds of data was non-speech segments, and it was thus omitted from transcription. This omission also resulted in slight increase of accuracy of the system as the non-speech parts were not transcribed into jibberish.

  The initial research introducing the main idea and basic smoothing model was presented in (Mateju, Cerva, and Zdansky, 2016) at SIGMAP 2016 held in Lisbon. The improved context-based smoothing model, which yields state-of-the-art results, was introduced in (Mateju, Cerva, Zdansky, and Malek, 2017) at ICASSP 2017 conference in New Orleans. Finally, an extended version of this work presenting more experiments and in detail evaluation on QUT-NOISE-TIMIT was published in (Mateju, Cerva, and Zdansky, 2017). The papers also reported the benefits of employing the proposed speech activity detection approach in conjunction with speech recognizer.

  The proposed speech activity detection approach follows all the set goals, and it is now considered as fully completed.

- Speaker Change Point Detection - **in progress**

  The speaker change point detection approach is in early stage of development. The current approach utilizes feed-forward deep neural network as classifier

and weighted finite state transducers (with forced duration of transition) as decoder to mark the transitions between two speakers. The network is trained on a mixture of real and artificially mixed broadcast data of Czech TV/radio shows. It is also possible to use it for online streams as it operates with low latency & real-time factor. The achieved results are promising, but they still leave much to be desired. Further research is thus required.

The further research will be focused on different input feature vectors as the log filter banks may not contain all the necessary information to distinguish between speakers. Features, such as i-vectors, d-vectors, etc., could be employed to improve the results. Another possibility, as the literature suggests, is to employ different, more complex, classifier, such as convolutional or long short term memory recurrent neural networks (bidirectional). It is also possible to further tune WFST transduction scheme (e.g. to distinguish between four different transitions (female-female, female-male, male-female, male-male)). A mixture of more complex features & classifier with WFST-based decoder should yield better results.

The plan is to get the first paper published in 2018. The final proposed speaker segmentation approach (fulfilling all the set goals) will be then integrated into TVR monitoring system developed at SpeechLab@TUL. It will provide the monitoring system with functionality of detecting speaker homogeneous segments.

References

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

Benyassine, A., E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit (1997). "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, pp. 64–73. ISSN: 0163-6804.

Bonastre, J., F. Wils, and S. Meignier (2005). "ALIZE, a Free Toolkit for Speaker Recognition". In: *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, pp. 737–740.

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

Bredin, H. (2017). "TristouNet: Triplet Loss for Speaker Turn Embedding". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 5430–5434.

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

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

Chen, K. and A. Salman (2011). "Learning Speaker-Specific Characteristics With a Deep Neural Architecture". In: *IEEE Trans. Neural Networks* 22.11, pp. 1744–1756.

Chen, S. S. and P. S. Gopalakrishnan (1998). "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion". In: *Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998*, pp. 127–132.

Chu, S. M., H. Tang, and T. S. Huang (2009). "Fishervoice and Semi-Supervised Speaker Clustering". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pp. 4089–4092.

Chung, H., S. J. Lee, and Y. Lee (2013). "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, pp. 700–703.

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

Dean, D., S. Sridharan, R. Vogt, and M. Mason (2010). "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*, pp. 3110–3113.

Desplanques, B., K. Demuynck, and J. Martens (2015). "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, pp. 3081–3085.

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

Evangelopoulos, G. and P. Maragos (2005). "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*, pp. 685–688.

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

Fergani, B., M. Davy, and A. Houacine (2008). "Speaker Diarization Using One-Class Support Vector Machines". In: *Speech Communication* 50.5, pp. 355–365.

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

Galibert, O. and J. Kahn (2013). "The First Official REPERE Evaluation". In: *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia, Marseille, France, August 22-23, 2013*, pp. 43–48.

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

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

Gao, C., G. Saikumar, S. Khanwalkar, A. Herscovici, A. Kumar, A. Srivastava, and P. Natarajan (2011). "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*, pp. 2637–2640.

Geiger, J. T., F. Wallhoff, and G. Rigoll (2010). "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*, pp. 2330–2333.

Ghaemmaghami, H., B. Baker, R. Vogt, and S. Sridharan (2010). "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*, pp. 3118–3121.

Ghaemmaghami, H., D. Dean, S. Kalantari, S. Sridharan, and C. Fookes (2015). "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*, pp. 2292–2296.

Gish, H., M. H. Siu, and R. Rohlicek (1991). "Segregation of Speakers for Speech Recognition and Speaker Identification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1991, 14-17 April 1991, Toronto, Ontario, Canada*, pp. 873–876.

Gupta, V. (2015). "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*, pp. 4420–4424.

Hrúz, M. and Z. Zajíc (2017). "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, LA, USA, March 5-9, 2017*, pp. 4945–4949.

Hughes, T. and K. Mierle (2013). "Recurrent Neural Networks for Voice Activity Detection". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pp. 7378–7382.

Jati, A. and P. Georgiou (2017). "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*, pp. 3567–3571.

Kneser, R. and H. Ney (1995). "Improved Backing-off for M-gram Language Modeling". In: *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pp. 181–184.

Kotnik, B., Z. Kacic, and B. Horvat (2001). "A Multiconditional Robust Front-End Feature Extraction with a Noise Reduction Procedure Based on Improved Spectral Subtraction Algorithm". In: *EUROSPEECH 2001 Scandinavia, 7th European*

Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001, pp. 197–200.

Larcher, A., J. Bonastre, B. G. B. Fauve, K. Lee, C. Lévy, H. Li, J. S. D. Mason, and J. Parfait (2013). "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*, pp. 2768–2772.

Li, C., X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu (2017). "Deep Speaker: an End-to-End Neural Speaker Embedding System". In: *CoRR* abs/1705.02304.

Li, J., B. Liu, R. Wang, and L. Dai (2004). "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*, pp. 61–64.

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

Lu, L. and H. Zhang (2002). "Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis". In: *Proceedings of the 10th ACM International Conference on Multimedia 2002, Juan les Pins, France, December 1-6, 2002*, pp. 602–610.

Lu, L. and H. Zhang (2005). "Unsupervised Speaker Segmentation and Tracking in Real-Time Audio Content Analysis". In: *Multimedia Syst.* 10.4, pp. 332–343.

Lu, L., H. Zhang, and H. Jiang (2002). "Content Analysis for Audio Classification and Segmentation". In: *IEEE Trans. Speech and Audio Processing* 10.7, pp. 504–516.

Ma, J. (2014). "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*, pp. 1558–1562.

Magrin-Chagnolleau, I., A. E. Rosenberg, and S. Parthasarathy (1999). "Detection of Target Speakers in Audio Databases". In: *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999*, pp. 821–824.

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

Markov, K. and S. Nakamura (2007). "Never-Ending Learning System for On-line Speaker Diarization". In: *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pp. 699–704.

Mateju, L., P. Cerva, and J. Zdansky (2015). "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 Mecha-*

tronics, ECMSM, 2015, Liberec, Czech Republic, June 22-24, 2015, pp. 184–187.

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

Mateju, L., P. Cerva, and J. Zdansky (2017). "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*, pp. 341–358.

Mateju, L., P. Cerva, J. Zdansky, and J. Malek (2017). "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, LA, USA, March 5-9, 2017*, pp. 5460–5464.

McLaren, M., L. Ferrer, D. Castán, and A. Lawson (2016). "The Speakers in the Wild (SITW) Speaker Recognition Database". In: *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pp. 818–822.

Meignier, S., J. Bonastre, and S. Igounet (2001). "E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing". In: *Odyssey*. ISCA, pp. 175–180.

Meignier, S. and T. Merlin (2010). "LIUM SpkDiarization: an Open Source Toolkit for Diarization". In: *in CMU SPUD Workshop*.

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

Moattar, M. H. and M. M. Homayounpour (2009). "A Simple but Efficient Real-Time Voice Activity Detection Algorithm". In: *17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, Scotland, UK, August 24-28, 2009*, pp. 2549–2553.

Moraru, D., S. Meignier, C. Fredouille, L. Besacier, and J. Bonastre (2004). "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*, pp. 373–376.

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

Ng, T., B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselý, and P. Matejka (2012). "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, pp. 1969–1972.

Povey, D., 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 (2011). "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.

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

Räsänen, O. J., U. K. Laine, and T. Altosaar (2009). "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*, pp. 1851–1854.

Rouvier, M., G. Dupuy, P. Gay, E. el Khoury, T. Merlin, and S. Meignier (2013). "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*, pp. 1477–1481.

Ryant, N., M. Liberman, and J. Yuan (2013). "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*, pp. 728–731.

Saon, G., S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury (2013). "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*, pp. 3497–3501.

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

Siegler, M. A., U. Jain, B. Raj, and R. M. Stern (1997). "Automatic Segmentation, Classification and Clustering of Broadcast News Audio". In: *Proc. DARPA Speech Recognition Workshop*, pp. 97–99.

Sinha, R., S. E. Tranter, M. J. F. Gales, and P. C. Woodland (2005). "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*, pp. 2437–2440.

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

Soldi, G., C. Beaugeant, and N. W. D. Evans (2015). "Adaptive and Online Speaker Diarization for Meeting Data". In: *23rd European Signal Processing Conference, EUSIPCO 2015, Nice, France, August 31 - September 4, 2015*, pp. 2112–2116.

Sriskandaraja, K., V. Sethu, P. N. Le, and E. Ambikairajah (2015). "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*, pp. 2297–2301.

Stern, R. (1997). “Specifications of the 1996 Hub-4 Broadcast News Evaluation”. In: *Proc. of the DARPA Speech Recognition Workshop*.

Thomas, S., G. Saon, M. V. Segbroeck, and S. S. Narayanan (2015). “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*, pp. 4500–4504.

Tranter, S., K. Yu, G. Evermann, and P. C. Woodland (2004). “Generating and Evaluating Segmentations for Automatic Speech Recognition of Conversational Telephone Speech”. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, pp. 753–756.

Tsiartas, A., T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. V. Segbroeck, A. Potamianos, and S. Narayanan (2013). “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*, pp. 718–722.

Vijayasenan, D. and F. Valente (2012). “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*, pp. 2170–2173.

Wang, Q., J. Du, X. Bao, Z. Wang, L. Dai, and C. Lee (2015). “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*, pp. 2282–2286.

Wang, R., M. Gu, L. Li, M. Xu, and T. F. Zheng (2017). “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, LA, USA, March 5-9, 2017*, pp. 5420–5424.

Wisdom, S., G. Okopal, L. E. Atlas, and J. W. Pitton (2015). “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*, pp. 4505–4509.

Yella, S. H. and A. Stolcke (2015). “A Comparison of Neural Network Feature Transforms for Speaker Diarization”. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 3026–3030.

Yin, R., H. Bredin, and C. Barras (2017). “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*, pp. 3827–3831.

Zhang, X. and D. Wang (2014). “Boosted Deep Neural Networks and Multi-Resolution Cochleagram Features for Voice Activity Detection”. In: *INTER-*

SPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pp. 1534–1538.

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

Zhu, W. and J. W. Pelecanos (2016). "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*, pp. 5045–5049.

Zibert, J., F. Mihelic, J. Martens, H. Meinedo, J. P. Neto, L. D. Fernández, C. García-Mateo, P. David, J. Zdánský, M. Pleva, A. Cizmar, A. Zgank, Z. Kacic, C. Teleki, and K. Vicsi (2005). "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, pp. 629–632.

Author's Publications

2017:

1. Mateju L., P. Cerva, and J. Zdansky (2017). "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*, pp. 341–358.

2. Safarik R. and L. Mateju (2017). "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, Proceedings*, pp. 402–410.

3. Mateju L., P. Cerva, J. Zdansky, and J. Malek (2017). "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, LA, USA, March 5-9, 2017*, pp. 5460–5464.

2016:

4. Bohac M., L. Mateju, M. Rott, and R. Safarik (2016). "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, Proceedings*, pp. 540–547.

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

6. Safarik R. and L. Mateju (2016). "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*, pp. 311–314.

2015:

7. Mateju L., P. Cerva, and J. Zdansky (2015). "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*, pp. 184–187.