February 2023

Master's Thesis, Master of Science in Engineering Acoustics

Developing a Robust Voice Activity Detection Algorithm of Individual Speakers in Speech-In-Speech Recordings

Alexander Kittel

Dept. of Health Technology, Hearing Systems Section, DTU

Supervisor: Tobias May

Associate Professor, Dept. of Health Technology, Hearing Systems Section, DTU

Supervisor: Eline Borch Petersen

PhD, Senior Scientist, WSAudiology

Abstract

Reliably detecting individual speaker activity in speech-in-speech recordings is vital for the analysis of conversational dynamics. Traditional voice activity detection (VAD) methods have been shown to perform well for speech-in-noise, but are not suitable when competing talkers are present. Manually annotated recordings of real conversations supplied by WSAudiology are combined with synthetically mixed speech and noise to form a training set for long short-term memory (LSTM) recurrent neural networks. These are trained and operate on mel-frequency cepstral coefficients (MFCC), and shown to outperform traditional VAD methods on unseen portions of the WSA dataset, as well as on unseen speech-in-noise mixtures. However, the trained classifiers fail to achieve good results in a cross-corpus test on recordings from the EasyCom dataset by Meta Research. Further work is required in order to properly examine the generalizability of this approach and better leverage the predictive power of neural networks for VAD of speech-in-speech.

1 Introduction

An emerging field in hearing science examines how listening conditions and hearing impairment affect the dynamics of a conversation[1–3]. Conversational interaction between two or more speakers is described based on on speech levels, the timing of turn-taking, duration of speaking and the articulation rate (speed of speaking). All of these measures require a robust and reliable detection of utterances made in the conversation by each individual speaker. Typical recording protocol of a dyadic or triadic conversation involves using directional cheek-mounted boom-arm microphones for each talker. Unfortunately, these may pick up interference from the other participants[4]. While detecting speech in general is not a problem, isolating the speech of each talker in their own recordings has proven difficult with conventional voice activity detection (VAD). First, the method must be robust towards the crosstalk from other interlocutors. Second, the voice detection must accurately include the onset and offset of the utterances so that the subsequent measurements of conversational dynamics are

valid. Third, in order to apply a learning-based approach, a corpus of recorded conversations with annotated speaker onsets and offsets for every participant is required for training, validation and testing.

At its simplest, VAD is a binary classification problem: '1' for speaker activity and '0' for everything else. Although historically developed to improve the intelligibility of noisy long-distance transmissions[5], VAD classifiers have become an important enabling technology for a multitude of uses[6–9]. VAD has evolved from simply thresholding signal features such as energy, zero-crossing rate, etc. to discriminate between speech and noise[10–12], to using adaptive thresholding for better performance on varying noise types and levels[13,14]. More complex features like spectral shape or harmonic structure with better discrimination between speech and background noise, along with statistical model-based approaches for the decision threshold[15] have since been shown to increase VAD performance in speech-in-noise applications[8,12,13] .

When the problem moves away from strictly discriminating speech from noise and into the realm of speech-in-speech recordings, even adaptive thresholds and otherwise highly discriminative features come up short[16]. The key difference arises from the interference (i.e. crosstalk from other speakers) now having very similar properties to the desired speech signal, making differentiation more difficult. As such, a new approach is required to discriminate between different speakers without compromising on speech-in-noise performance. A widely used concept allowing for the efficient extraction of speech-significant signal information is the mel-frequency cepstral coefficients (MFCCs)[17–24]. The motivation for using this signal representation in the context of VAD is its approximate similarity to human sound production: the time-frequency transform of a logarithmic spectrum contains information about the harmonic structure caused by the fundamental frequency of speech as well as larger envelope-like structures corresponding to vocal tract resonances (also known as speech formants)[25].

Recently, the advent of machine learning and deep neural networks (DNN) has enabled VAD to be treated as a supervised learning problem, with the added benefit of recurrent methods like long short-term memory (LSTM)[26]. This method enables access to temporal context, which has been shown to play a critical role in both human and network-based speech processing[27]. Another major advantage of using a neural network to classify speech is the relative ease of employing multidimensional signal features like MFCCs.

Using labeled recordings of triadic conversations from a recent WSAudiology study on the effects of amplification and directionality in hearing aids on conversational dynamics[4], this paper will present the development and evaluation of a learning-based approach to speech-in-speech VAD. The goal of this thesis is to demonstrate that a VAD network trained on sequences of MFCCs is capable of accurately detecting the voice activity of a target speaker in a conversation. Specifically, performance should exceed that of existing VAD techniques without compromising on performance in classical speech-in-noise scenarios. The generalizability of a trained network is examined by subjecting it to cross-corpus testing on unseen recordings. Ideally, the high predictive power should not come at the expense of overfitting.

Two established voice activity detectors developed by Ghosh et al.[11] and Graf et al.[13] have been selected for baseline comparative analysis and are presented in Section 2.3. Performance on recordings from real conversations is evaluated on the data from WSAudiology[4] (referred to as the WSA dataset), as well as on recordings from the EasyCom Augmented Reality dataset by Meta Research[28] (referred to as the EC dataset). Additionally, robustness to more typical speech-in-noise testing is done using synthetic speech and noise mixtures of TIMIT[29] sentences and DEMAND[30] noise. The VAD performance is analyzed in relation to each dataset and results are presented in Section 5. Observations are summarized and interpreted in Section 6.

2 METHODS

Although training a DNN is technically possible on raw audio data, it involves very high computational complexity and much deeper networks than considered in this paper[31]. Extracting more descriptive signal representations such as MFCCs not only allows for less complex networks, it also reduces the size of the feature space used for training. This makes it possible to develop and train a VAD network with feasible computing times and loads.

2.1 Mel-frequency cepstral coefficients

The formants and fundamental frequency (F0) of speech are two of the most pertinent characteristics used for phonetic analysis. According to the source-filter theory of speech[32], these describe the process of human sound production: the F0 is the frequency of excitation in the vocal folds, and the formants are determined by the resonant frequencies of the vocal tract[33]. It would therefore be beneficial to emphasize these characteristics in a feature space designed to be discriminative of different speakers as well as between speech and non-speech[19]. One such adapted feature space is the MFCC representation, which makes use of several techniques to capture F0 and formants while also avoiding informational redundancy[20]. This section will go over the procedure for extracting MFCCs from an audio signal for use in machine learning and VAD.

The signal is initially windowed and the discrete Fourier transform (DFT) is calculated for each window. The features we are interested in representing can be understood as harmonic components of the frequency representation of speech. By applying another time-frequency transform to the log-spectrum, we can model them even clearer. The resulting inverse spectrum representation is known as the signal *cepstrum* and was pioneered by Bogert et al.[34].

The full signal cepstrum still includes excessive information for speech modeling, considering the fundamental frequency and first formants of speech are relatively low-frequency[32]. As such, we introduce a step in the algorithm before taking the log-magnitude of the DFT: By mapping the spectrum onto a perceptually motivated mel scale filterbank (composed of overlapping triangular filters of increasing bandwidth)[20,35], the amount of frequency information

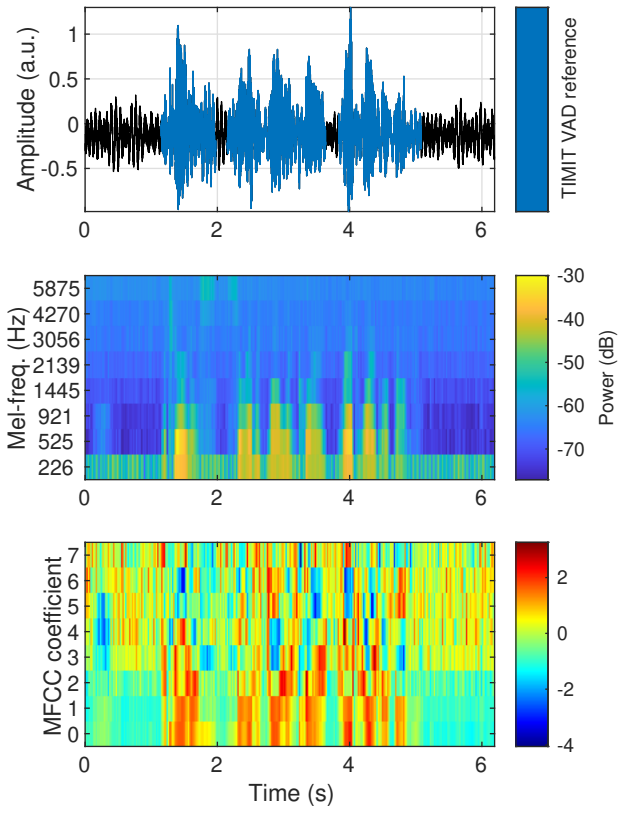

Figure 1. Top: TIMIT "*materials ceramic modeling clay red white or buff*" test sentence mixed with DEMAND kitchen noise at 0 dB SNR. Middle: 8-band mel frequency spectrum. Bottom: normalized 8 MFCC representation.

is dramatically reduced without compromising the low-frequency region of interest. The following step of taking the log-magnitude ensures we have perceptual representations of both the magnitude and frequency scales[20] (middle plot of Fig. 1).

Due to overlapping filters in the filterbank, neighboring samples in the log-mel-spectrum are highly correlated. In order to avoid redundant information and decorrelate the final coefficients, Davis and Mermelstein[36] introduced the use of the discrete cosine transform (DCT) as the final time-frequency transform in the MFCC extraction. As shown in the bottom plot of Fig. 1, this signal representation is not easy to interpret visually.

The full MFCC algorithm is summarized as follows:

1) Apply DFT to the windowed signal segments to get the spectra.
2) Apply the mel-filterbank to the spectra and sum the energy within each band.
3) Take the logarithm of the filterbank energies.
4) Take the DCT of the log filterbank energies.

2.2 LSTM architecture

Recurrent neural networks using long short-term memory (LSTM) layers[37] have been shown to produce good results in speech classification on a variety of test data[17,26]. LSTM layers benefit from the ability to consider the temporal context of data sequences: either by using information from only previous samples, or by processing data in both directions and utilizing future context as well. This is known as bidirectional LSTM, or biLSTM[18]. This section will briefly explain LSTM functionality from a conceptual point of view and describe the general architecture of the networks considered in this paper.

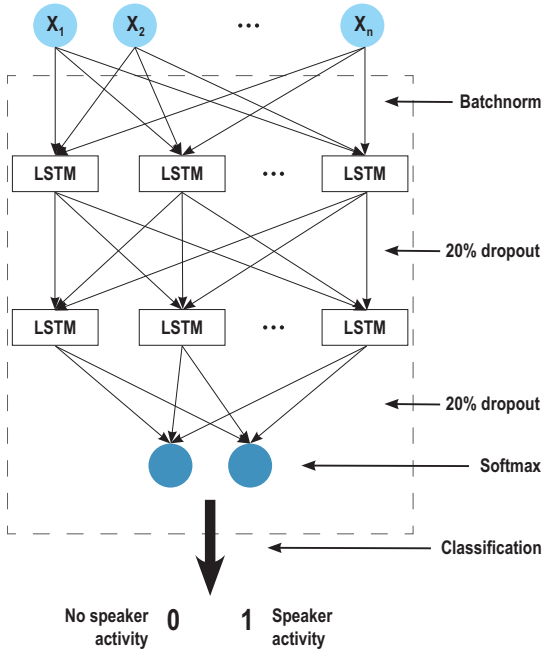

Figure 2. Two layer LSTM network structure. Each row of LSTM cells comprises a single LSTM layer. Extracting the features used in the input vectors $X_1, X_2 \dots X_n$ is done upstream of the network. The final output is a binary frame-based vector indicating speaker activity.

LSTM cells, or neurons, are able to learn long-term dependencies in data sequences thanks to their internal architecture. The state (stored memory) of each cell in a layer is decided by a set of gates: an *input gate*, a *forget gate* and an *output gate*. These gates regulate how information is kept or discarded using sigmoid and hyperbolic tangent activation functions.

The first step in this process is deciding what portion of the information already stored is to be kept and what portion is to be discarded. This is done at the forget gate by comparing new information with memory. Next, the input gate determines what portion of new information should be remembered. The output gate performs the final step: Deciding what information is relevant to output from the neuron to the next layer in the network. biLSTM cells work by letting the input data pass in both directions, which allows the cell to memorize both previous and subsequent events[38].

The neural networks examined in this paper follow the general structure presented in Fig. 2. The task of the network is sequence classification, assigning either a '0' for no speaker activity or a '1' for speaker activity to each frame of a signal. The inputs $X_1, X_2 \dots X_n$ denote vectors of signal features; in the case of an 8-coefficient MFCC representation

of a 5-minute recording sampled at 16 kHz using 15 ms analysis windows with 50% overlap, the input would be $n = 8$ vectors of roughly 40000 coefficients. Accordingly, the output would be a single binary vector with the same length.

Following recommendations from Ioffe and Szegedy[39] and Srivastava et al.[40], the input training features are normalized in batches in a process designated *batchnorm*[39], before being passed to the first LSTM layer. Regularization is performed by randomly setting 20% of the outputs to 0 using *dropout*. This helps prevent overfitting and should increase VAD performance on unseen data[40]. A two-cell *softmax* layer combines the last LSTM layer outputs and assigns a speaker activity probability between 0 and 1 to each signal frame[41]. During training, a cross-entropy loss function is used to penalize this probability depending on how far from the reference label it is[42]. The resulting weights and biases are used in the final binary classification layer. The effects of batch-normalization and dropout layers are investigated as part of the hyperparameter validation portion of Appendix A.

2.3 Baseline VAD techniques for comparison

The performance of our trained classifiers is compared to two well-documented VAD techniques designed by Ghosh et al.[11] and Graf et al.[13]. These will act as performance baselines. This section will describe the fundamental features utilized by these two methods, as well as the values of specific parameters used moving forward. Parameters shared across methods such as analysis window type, duration and overlap are kept the same across all methods and will be described in Section 3.

2.3.1 Long-term signal variability

The first baseline VAD method was presented by Ghosh et al.[11] and relies on a measure of long-term signal variability (LTSV). By relying on the nonstationary properties of speech, the variability of noisy speech is different to that of only noise. This technique considers signal variability in segments of 300 ms rather than the typical 10 to 20 ms window duration of short-time analysis[11].

As the LTSV changes depending on whether speech is present in a noisy signal, a threshold is applied. By using a convex combination of previous LTSV measures labeled either noise or noisy speech, a single parameter α is used to update the threshold; a high value for α sets it closer to the expected value for noisy speech. In their testing, Ghosh et al.[11] show LSTV to be a good discriminator for VAD, especially in noisy conditions (-10 to -5 dB SNR). The parameters specific to this method are set according to the best performance achieved in the original paper: the LTSV is calculated across 300 ms segments on a Bartlett-Welch estimated spectrum[43] using 10 ms averaging. The convex combination parameter is set to $\alpha = 0.6$.

2.3.2 Modulation-phase difference

Graf et al.[13] proposed a VAD scheme for use in in-car communication (ICC) systems based on a combination of two modulation-based features. The first modulation feature

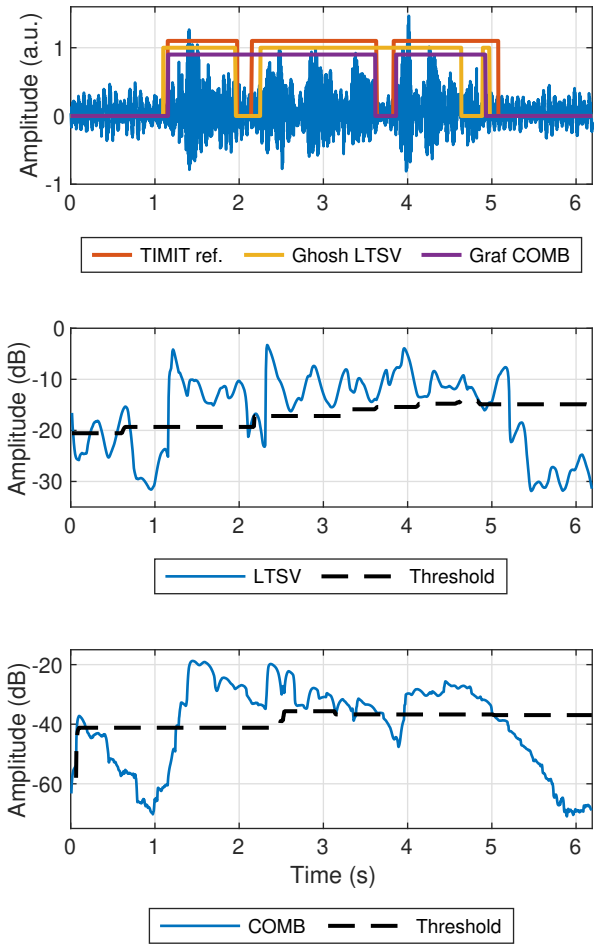

Figure 3. Top: Reference, LSTV and COMB VAD outputs on TIMIT "*materials ceramic modeling clay red white or buff*" test sentence mixed with DEMAND kitchen noise at 0 dB SNR. Middle: Ghosh et al.[11] LTSV feature and threshold. Bottom: Graf et al.[13] COMB feature and threshold.

(MOD) is based on the syllable rate of typical human speech. It is observable in a spectrogram as a modulation of roughly 4 Hz. The second feature is the modulation-phase difference (MPD), that assumes an alternating structure of voiced (low-frequency) and unvoiced (high frequency) phonemes. A hold scheme is implemented to temporally extend the MPD detections for a set amount of time. The combined feature (COMB) is achieved by multiplying MOD and MPD.

Using COMB is shown to be beneficial for VAD in ICC, as the feature extraction has low computational complexity and performs well with low frequency resolution[13]. This VAD method is implemented in this work using the same adaptive thresholding procedure proposed by Ghosh et al.[11], with the same α-parameter value of 0.6. For this method, modulation differences are expected to happen in two subbands: 200 to 2000 Hz and 4500 to 8000 Hz. The hold scheme duration is set to 300 ms.

3 Experiments

In this section, the experimental procedure used for testing the performance of the trained VAD networks will be summarized. This includes the selected feature extraction

parameters and network selection, as well as the setup utilized for simulating speech-in-noise and the recording protocols of the two real conversation datasets used in this paper.

3.1 Network and feature parameters

Following the validation procedure outlined in Appendix A, a base network configuration for the trained classifiers is chosen, including both dropout-based regularization and batch-normalization. The input training sequences are 200 frames (equivalent to 3 seconds of signal for a window size of 15 ms) and each LSTM layer contains 100 hidden units. Two networks are trained on 8 MFCCs using LSTM and biLSTM, respectively. The third network is trained on the mel-frequency spectrum using biLSTM to see if the final DCT computation can be spared. The spectrum has 8 bands in the filterbank, to keep the feature-spaces equally sized across networks. The training set consists of 80% WSA training-reserved data and 20% speech-in-noise mixtures of TIMIT training sentences and DEMAND cafeteria noise, on a range of -20 to 20 dB SNR. The complexity of the networks is in the order of 340 thousand learnable parameters. All features (including the features required for the Ghosh et al.[11] and Graf et al.[13] VAD schemes) are extracted using 15 ms[44] Hann analysis windows with 50% overlap.

3.2 Real recording datasets

3.2.1 WSA dataset

This paper is centered around the WSA dataset, as recorded by Borch Petersen[4]. The set contains recordings of a total of 75 people spread over 25 triads. These were made up of one hearing-impaired and two normal-hearing test subjects recorded over 12 trials of 5 minute conversation. The experiment was conducted in a typical office-like room. The background noise was recorded locally in the WSA canteen and played back from three speakers at two levels, depending on the scenario: *quiet* at 50 dB SPL and *noisy* at 75 dB SPL. Participants were seated one meter from each other around a table and recorded individually using cheek-mounted boom-arm microphones. The dataset totals 25 hours of conversation, equating to 75 hours of microphone recordings. The two noise-level scenarios will not be considered different for the purpose of this work.

Speaker activity was labeled manually by an external service. Following typical speech pause and gap durations presented by Heldner and Edlund[45], gaps less than 180 ms are bridged, and utterances shorter than 90 ms are removed as part of the labeling process. The full dataset is partitioned along triads into 60% training, 10% validation and 30% testing sets. This ensures no participant is present in more than one subset. The label distribution is on average 31% main speaker activity per recording. This is coherent with three speakers per conversation and an average of 93% speech presence per trial.

3.2.2 EC dataset

In order to gauge the generalizability of the trained classifiers, performance will also be measured on the EasyCom dataset made available by Donley et al.[28]. Similarly to the WSA data, test subjects were recorded individually while participating in group conversation. The recording environment was a mildly reverberant room (RT60 of 645 ms) with 10 loudspeakers playing spatially uncorrelated restaurant-like noise fixed at 71 dB SPL. The number of people in each conversation did vary, but at least two were equipped with cheek-mounted microphones. Consequentially, not every speaker in the conversations was individually recorded and labeled.

Since the intended use of the dataset is towards augmented reality (AR), it contains video and audio from microphones mounted on special AR-glasses. Recorded audio from the glasses will not be considered in this work. Further, any file without speech activity from the microphone-wearers is removed, leaving roughly 11 hours of usable recordings. As with the WSA data, speaker activity labeling was performed manually. The data is only used for testing the classifiers, and is therefore not split into any subsets. The distribution of speech to non-speech per recording is slightly lower than in the WSA data. On average, each file has 26% main speaker activity, though this number does vary greatly across files. The total proportion of speech to background noise is not available, as not all participants in the conversations are labeled.

3.3 Synthetic speech-in-noise data

Due to the scarcity of properly labeled real conversational recordings, most VAD methods are evaluated on synthetic mixtures of speech and noise[11,13,14,24,26,46,47]. The DEMAND database[30] was developed to supply multi-channel recordings of background noise in a variety of indoor and outdoor environments. Five of the in-total 18 noise-types are used in this paper: *restaurant*, *cafeteria*, *office*, *kitchen* and *living room*. The cafeteria noise is reserved for the speech-in-noise add-in to the WSA training data, and the restaurant noise is reserved as add-in for the validation set. As such, testing is conducted on mixtures using the remaining noise-types.

Clean speech is taken from the TIMIT database[29], which consists of a large number of recorded short sentences in English (192 female and 438 male speakers). Mixtures matching the five-minute duration of the WSA recordings are created by chaining and equally spacing 5 to 10 TIMIT sentences into a single file before superimposing DEMAND noise scaled to specific SNRs. These are substantially more speech-sparse compared to the real recording data: roughly 15% speech activity per file, with no competing talkers. Each mixture is designed with a leading 0.5 s of noise only. The full test set totals 1350 mixtures (roughly 112 hours) sampled at 16 kHz. Reference labels are generated by combining the included TIMIT time-aligned phonetic transcriptions of each sentence. The TIMIT corpus comes pre-divided into two subsets, making it simple to keep training and testing disjoint.

4 Evaluation

4.1 Receiver operating characteristics

The performance of a VAD scheme is usually expressed in terms of receiver operating characteristic (ROC) statistics.

These compare the reference class of a signal frame with the classification output of a model[48,49]. The outcome of a decision by a classifier can be described in any one of four ways:

- True Positive (TP): correct classification of a speech frame
- False Negative (FN): incorrect classification of a speech frame as a non-speech frame
- True Negative (TN): correct classification of a non-speech frame as such
- False Positive (FP): incorrect classification of a non-speech frame as a speech frame

From these, many metrics can be calculated, chief among which are the true positive rate (also called hit-rate and denoted TPR), false positive rate (also called false alarm rate or FPR), false negative rate (miss-rate or FNR) as well as the hit-rate minus false alarm rate (HFA) and the overall accuracy (ACC). For a given signal with p speech-labeled and n non-speech labeled frames, these metrics are given by:

$$\begin{aligned}
\mathrm{TPR} &= \frac{\mathrm{TP}}{p} \\
\mathrm{FPR} &= \frac{\mathrm{FP}}{n} \\
\mathrm{FNR} &= \frac{\mathrm{FN}}{p} \\
\mathrm{HFA} &= \mathrm{TPR} - \mathrm{FPR} \\
\mathrm{ACC} &= \frac{\mathrm{TP} + \mathrm{TN}}{p + n}
\end{aligned}$$

4.1.1 *Crosstalk error-rate*

An error estimation specific to overlapping speech from competing talkers is warranted for speech-in-speech VAD performance analysis. There is not much existing work on this, but the availability of annotated speaker activity for all participants makes a simple measure possible. The hit-rate of the VAD on the combined labels of all competing talkers in a conversation is in essence an error-rate specific to crosstalk. This measure does depend on the labeling protocol of a given dataset, but provides a simple and useful pointer towards what type of error a classifier may tend to make. The metric should not be used in a vacuum however, as the ROC context is very important for interpretation.

4.2 Dynamic utterance measures

The goal of the trained VAD networks is to enable the study of conversational dynamics. As such, more dynamic metrics are required to properly evaluate their behavior. Freeman et al.[50] proposed a set of measures that would describe the performance of voice activity classification from a more speech dynamic perspective, by evaluating metrics based around utterances instead of averaged across the entire signal duration. In this way, the VAD reaction to speech onsets and offsets may be captured.

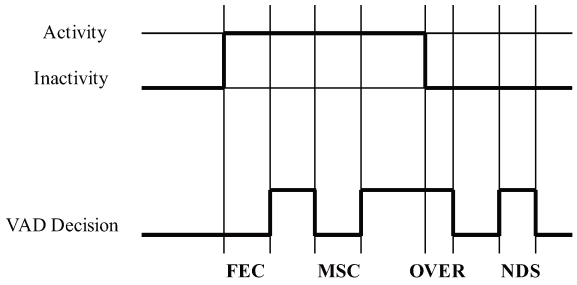

Figure 5. Example of the four intervals covered by the dynamic utterance measures[50], as presented by Beritelli et al.[51]. The metrics are measured using the reference utterance onsets and offsets, and the detection start and end points.

The measures proposed have since been used extensively to analyze VAD performance[11,16,47,52,53], and include four error-types (see Fig. 5):

- clipping at the front of an utterance (FEC)
- clipping in the middle of an utterance (MSC)
- noise (or interference) detected as speech (NDS)
- VAD hangover time after an utterance (OVER)

As with ROC statistics, these metrics are calculated with regards to the reference classification of each signal frame. FEC is calculated as the time-lag between the reference onset of an utterance and the first following positive detection. Similarly, OVER is calculated as the time-lag between the offset of an utterance and the next detection stop. Following the practice used by Graf et al.[47], MSC and NDS are expressed as the rate of incorrect detections in regions not covered by FEC and OVER.

5 Results

5.1 VAD detection on conversation datasets

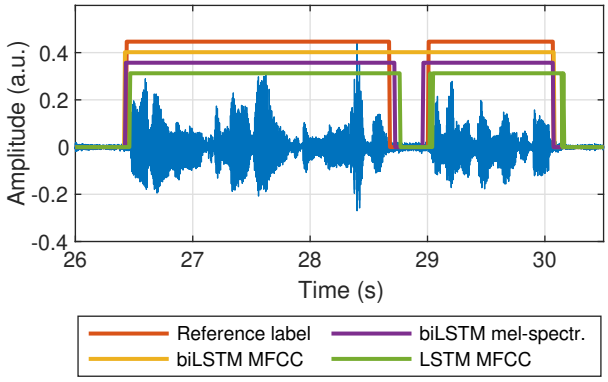

Figure 6. Example of VAD outputs and reference speaker activity for a section of a recording from the WSA test data. The main speaker is Talker 1 of triad 1, in their 11th trial saying (in Danish) "*så var der en af mine kollegaer som–øh, altså, han var lige bleven [sic] ansat*". Note the biLSTM MFCC network is erroneously labeling the pause before the 29 second mark as speech. The LSTM MFCC network seems to have a longer delay before detection in front of the utterances, as well as more overhang after.

VAD performance metrics on both the WSA test data and the EC dataset are presented in Fig. 4. The goal of this study is first and foremost to show trained networks can achieve

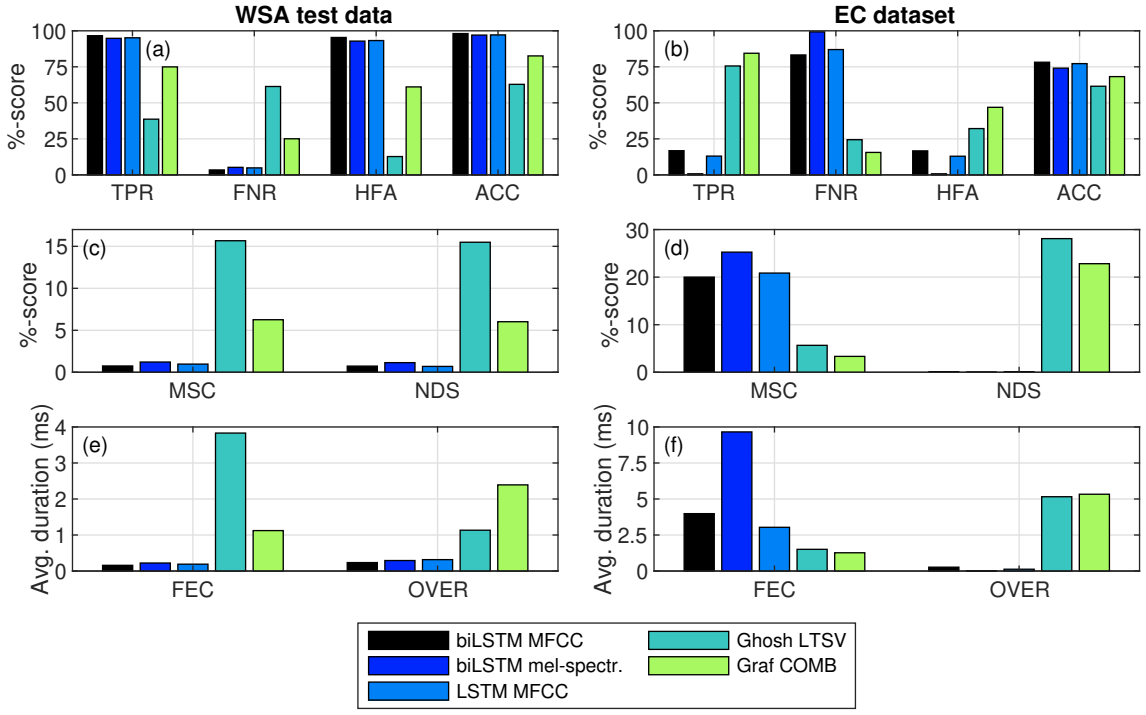

Figure 4. VAD performance metrics on real recording datasets. On the left: WSA test data. On the right: EC data. (a,b): hit-rate (TPR), miss-rate (FNR), hit-rate minus false alarm (HFA) and accuracy (ACC). (c,d): clipping rate in the middle of utterances (MSC) and rate of interference detected as speech (NDS). (e,f): averaged front-of-utterance detection lag (FEC) and overhang (OVER). Note the y-axes are not scaled equally across datasets.

precise utterance detection on the WSA data, and it is clear there is a net benefit to the learning-based approaches: The trained classifiers all achieve HFA rates above 93% and an ACC of 97% (Fig. 4.a). MSC and NDS both lie around 1%, with FEC and OVER well below 0.5 ms (Fig. 4.c and Fig. 4.e). The dynamic measures indicate the trained methods capture each utterance with a high degree of precision (very low lag in front and very low overhang after). There may be a slight advantage towards the biLSTM MFCC combination when considering all results, but this is very small. Both baseline VAD schemes perform markedly worse on this data, especially the LTSV-based scheme. The LTSV feature itself is ill-suited for differentiating between speakers, as the signal variability is not markedly different between main speaker activity and crosstalk dominated sections. This is illustrated by the very low HFA of 13%, as well as the high NDS of 15% in Fig. 4.c and Fig. 4.e.

Comparing metrics across datasets in Fig. 4, it is clear performance on the EC data is not as good as on the WSA test data (Fig. 4.b, Fig. 4.d and Fig. 4.f). Low HFA (below 25%) and high MSC (20% to 25%) indicate the networks are heavily mislabeling within utterances. To corroborate this, TPRs are very low (17%, 0% and 13%) and OVER and NDS are very close to 0 (around 0.5 ms and 0.1% respectively). High FEC (3 to 10 ms) indicates they are not tracking onsets very well on this data. Both baseline VADs struggle on this data as well, though seemingly in a very different way to the networks. Very high NDS (well above 20 ms) and low MSC (around 3% and 6%) points towards a similar difficulty as for the WSA test data: mislabeling primarily happens outside of utterances. The FEC is below 2.5 ms for both baseline schemes, which is an indicator of good onset tracking. It should be noted the dynamic measures do not account for completely undetected utterances, and may be misleading if considered without the context of the HFA.

5.1.1 *Crosstalk-specific error rate*

Table 1
Crosstalk error-rate measure: hit-rate on the main talker label subtracted from competing talkers. Note: due to recording differences, these results are not directly comparable across datasets.

VAD	WSA test data	EC dataset
biLSTM MFCC	14.5%	0.15%
biLSTM mel-spectr.	14.6%	0.01%
LSTM MFCC	14.6%	0.11%
Ghosh scheme	28.0%	48.4%
Graf scheme	22.0%	42.8%

Table 1 shows the 'crosstalk error -rate' for the WSA test data and the EC dataset as described in Section 4.1.1. Due to the fact that not every participant in the EC conversations has a corresponding labeled recording of themselves, this crosstalk metric cannot encompass all interfering speech. All participants are recorded and labeled in the WSA recordings, meaning the metric is more indicative of actual VAD behavior on that dataset. Due to this discrepancy, the crosstalk metric is not comparable across datasets. However, the crosstalk error-rate does still point towards trained classifiers being more robust against interfering speech than the baseline VAD schemes for the WSA test set. Another point of note is that the metric will be 0 or close to 0

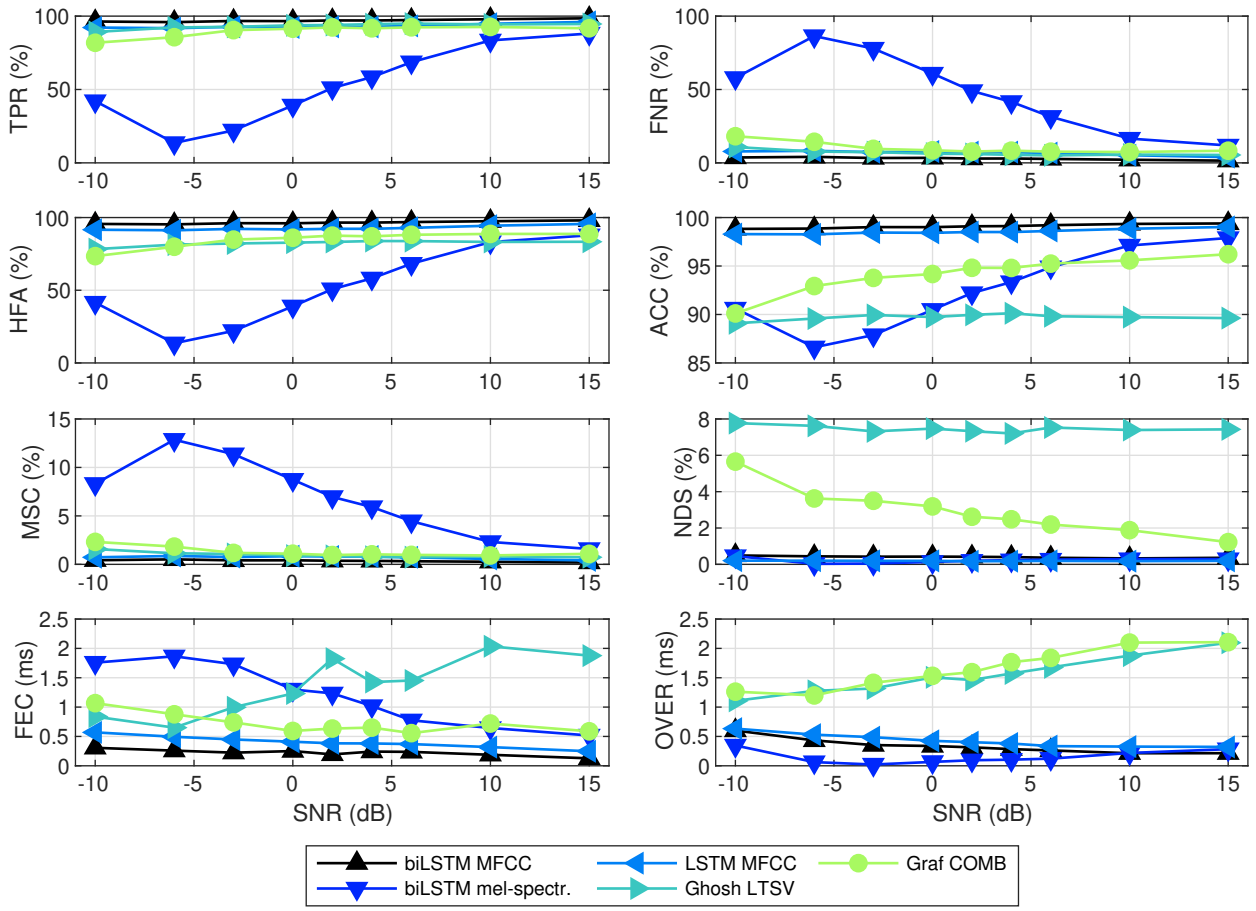

Figure 7. VAD performance metrics on synthetic speech-in-noise mixtures as a function of SNR. On the left side, from the top: true positive rate (TPR), hit-rate minus false alarm (HFA), clipping rate in the middle of utterances (MSC) and front-of-utterance time lag (FEC). On the right: true negative rate (TNR), accuracy (ACC), rate of noise detected as speech (NDS) and utterance-averaged overhang after speech (OVER). Note the y-axes are not balanced.

if the VAD method in question is under-labeling severely. Considering the very low HFA of the networks on the EC data in Fig. 4.b, the low crosstalk error-rate is more likely due to excessively conservative VAD on this data than to an inherent robustness to crosstalk.

5.2 VAD detection on speech-in-noise

Results from testing on the more typical speech-in-noise mixture data are presented for all VAD techniques in Fig. 7. The margin of improvement between the network-based approaches and the baseline VAD schemes is much smaller than on the WSA test set. Even so, it is notable that the trained classifiers still perform very well on speech-in-noise. From these results, it seems training on the mel-spectrum yields worse results for SNRs below 5 dB. This is apparent especially in the HFA rate, where a noticeable dip from 41% to 13% occurs from -10 to -6 dB SNR. The same behavior is illustrated in the MSC, TPR and FNR rates, but not in NDS, which suggests the errors are happening mostly inside of utterances. The networks trained on MFCCs show very good results for all metrics across the full range of SNRs. FEC and OVER are kept below 2 ms, which is substantially higher than on the WSA test set (Fig. 4.e), but still better than the baseline VAD schemes. The MFCC networks average 96% and 93% HFA for the biLSTM and LSTM respectively. As for the WSA test set, there seems to be a slight advantage to the biLSTM network, but differences are marginal. The baseline VAD schemes show somewhat worse performance in terms of NDS, FEC and OVER, which indicates they are worse at capturing sentence onsets and offsets, and mislabel noise more often. The LTSV-based scheme actually performs worse on FEC as the SNR improves, likely due to its reliance on noise being more stationary than speech. [1]

6 Discussion

6.1 LSTM VAD performance

The LSTM networks outperform the baseline VAD schemes on the WSA test set across all metrics. The drastic improvement in MS and, NDS in Fig. 4.c illustrates the learning-based approach is able to differentiate between main speaker activity and interference with a very low degree of error. Much lower FEC and OVER (Fig. 4.e) mean the networks have low detection lag and overhang, indicating higher precision in capturing utterance onsets and offsets than the baseline methods. This must all be seen in the light of very high TPR and HFA in Fig. 4.a. High hit-rate and low false-alarm rate mean the networks capture the main speaker activity when it happens. The crosstalk

1. The noise-types used in the test set are all non-stationary.

error-rate for this dataset (column one, Table 1) is able to include all participants in the WSA recordings, and shows the learning-based approaches are much more robust to interfering speech than the baseline methods.

Understanding the metrics in context is important especially when considering the performance on the EC dataset. Here, the networks have lower TPR and higher FNR (Fig. 4.b), meaning they hardly ever classify speaker activity correctly. Conversely, the baseline methods have very high TPR and low FNR which at face value seems to be a good thing. However, the low HFA in Fig. 4.b indicates a large number of false alarms. In essence, the networks are annotating too conservatively (not estimating enough speaker activity) and the baseline VADs are annotating too liberally (estimating too much speaker activity). This largely negates analyzing utterance lag and overhang: If there are no or very few detected utterances, FEC and OVER have no meaning. The validity of the crosstalk error-rate (column two in Table 1) is similarly affected: It will be close to zero if there is close to no detected speaker activity.

6.2 The challenge of a generalized solution to VAD in real conversations

The downside of using neural networks is the necessity for large volumes of data with organized labels, and the risk of trading high prediction power for low generalizability. It was not possible to demonstrate acceptable VAD on the EC dataset, either due to overfitting to the training data or because the unseen dataset is too difficult to classify. Worse recording quality, more sparse and quiet talking, different conditions etc. could all be plausible reasons. By the nature of machine-learning, it is highly speculative as to why exactly a network performs well on one dataset and poorly on another. One way to ascertain this would be to increase the number of datasets in testing, and attempt to understand which differences in the data lead to better or worse VAD. Unfortunately, manually annotating speech is a time-consuming and expensive process, meaning finding more datasets to establish a better picture of cross-corpus performance is difficult. If more granular labeling is desired, such as tone of voice, laughter, non-speech vocalizations etc. for conversational dynamics analysis, the same problem arises: there are not enough datasets with the proper annotation at the moment.

It might be enough to develop networks specialized to a single recording protocol using a kind of 'minimum training size', if the goal is just efficient and cheap labeling of conversations. If only a small portion (say 10%) of the data needs to be labeled manually in order to train a VAD network to satisfactorily annotate the remainder of the data, there is still a net benefit to be had. This would take advantage of the inherently high predictive power of neural networks without the worry of cross-corpus generalizability.

6.3 Further work

To circumvent the aforementioned issue with scarce data, it could be beneficial to investigate data augmentation in order to increase variance within a single dataset. Recordings can be randomly pitch- and time-shifted, extra noise can be injected, all to effectively increase the volume of a dataset with directly derived synthetic data [54].

Another avenue for further exploration could be the creation of an entirely simulated speech-in-speech dataset using speech and noise mixtures, taking into account things like the Lombard reflex [55] and the rate of pauses and gaps [45]. This would require an extensive model of conversational dynamics, but would benefit from total control over SNR, reverberation (room impulse responses could be used to simulate realistic conditions) and speaker overlap. There exist many databases of annotated short sentences similar to TIMIT [29] in multiple languages, as well as a multitude of background noise recording databases like DEMAND [30], making a large and highly varied dataset possible.

7 Conclusion

This paper presents a LSTM-based approach to VAD using neural networks, specifically intended for use in the analysis of conversational dynamics [1–3,45]. The aim of this research was to show better VAD performance is attainable for speech-in-speech when applying a learning-based method. This was motivated by the recent creation of an annotated corpus of recorded conversations for WSAudiology by Borch Petersen [4]. Classical VAD approaches by Ghosh et al. [11] and Graf et al. [13] are shown to have problems on this data, specifically in terms of crosstalk (see Table 1). Networks were trained on features extracted from a mixture of real recordings from the WSA training set and speech-in-noise mixtures of varying SNRs. Using a learning-based approach for VAD produced precise utterance labeling on the WSA test set with relatively low feature dimensionality. The networks greatly outperform the two classical VAD methods on the WSA test data, and marginally outperform them on typical TIMIT [29] and DEMAND [30] speech-in-noise mixtures. The experiments conducted illustrate the benefits of using LSTM-based networks for VAD in communication studies, by supporting ROC analysis with utterance-based metrics and a novel crosstalk error measure. Unfortunately, the networks did not perform well in the cross-corpus test using the EasyCom dataset by Meta Research [28]. Further work is required to ascertain exactly why that is the case, and how to increase generalizability of this LSTM-based VAD method trained on real conversational recordings.

Acknowledgments

I would like to thank my supervisor Tobias May for the opportunity to collaborate with him on multiple occasions during my master's program at DTU, and for his uncompromising guidance throughout this project. I would equally like to thank my supervisor Eline Borch Petersen for initially bringing the opportunity to conduct this research to the table, and for her insight into the intricacies of hearing science and audiology.
Special thanks to Marta Monrós for her love and encouragement, Janet Kino for her zealous proofreading of endless sections after sections and to all those that have helped by reviewing and providing comments.
I am extremely grateful for the unwavering support I've received from friends and family. It would have been a tiresome and lonely journey without them.

REFERENCES

[1] E. Borch Petersen, E. N. MacDonald, and A. Josefine Munch Sørensen, "The effects of hearing-aid amplification and noise on conversational dynamics between normal-hearing and hearing-impaired talkers," *Trends in Hearing*, vol. 26, pp. 1–18, 2022.

[2] A. J. M. Sørensen, E. N. MacDonald, and T. Lunner, "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," *International Symposium on Auditory and Audiological Research Isaar*, pp. 37–44, 2020.

[3] L. V. Hadley and J. F. Culling, "Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions," *Frontiers in Psychology*, vol. 13, p. 1061582, 2022.

[4] E. Borch Petersen, "Scientific Study Protocol, Effects of amplification and directionality on the conversational dynamics in triads," *Internal WSA report: unpublished*, 2022.

[5] J. M. Fraser, D. B. Bullock, and N. G. Long, "Over-All Characteristics of a TASI System," *Bell System Technical Journal*, vol. 41, no. 4, pp. 1439–1454, 1962.

[6] H. M. Chang, "'CrossTalk': Technical challenge to VAD-like applications in mixed landline and mobile environments," *IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, IVTTA*, pp. 77–80, 1996.

[7] N. Shankar, A. Kucuk, C. K. A. Reddy, G. S. Bhat, and I. M. Panahi, "Influence of MVDR beamformer on a Speech Enhancement based Smartphone application for Hearing Aids," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, vol. 176, no. 1. IEEE, jul 2018, pp. 417–420.

[8] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. Lopez Moreno, "Personal VAD: Speaker-Conditioned Voice Activity Detection," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 433–439.

[9] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, no. 1, pp. 7229–7233, 2013.

[10] J. Saunders, "Real-time discrimination of broadcast speech/music," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 993–996.

[11] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[12] E. Chuangsuwanich and J. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2645–2648, aug 2011.

[13] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Voice activity detection based on modulation-phase differences," *Speech Communication - 12. ITG-Fachtagung Sprachkommunikation*, pp. 80–84, 2016.

[14] X. Li, R. Horaud, L. Girin, and S. Gannot, "Voice activity detection based on statistical likelihood ratio with adaptive thresholding," *2016 International Workshop on Acoustic Signal Enhancement, IWAENC 2016*, 2016.

[15] J. Sohn, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[16] P. K. Ghosh, A. Tsiartas, P. Georgiou, and S. S. Narayanan, "Robust voice activity detection in stereo recording with crosstalk," *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, pp. 3098–3101, sep 2010.

[17] P. Sertsi, S. Boonkla, V. Chunwijitra, N. Kurpukdee, and C. Wutiwiwatchai, "Robust voice activity detection based on LSTM recurrent neural networks and modulation spectrum," *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, vol. 2018-Febru, pp. 342–346, dec 2018.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[19] S. Yadav and A. Rai, "Learning discriminative features for speaker identification and verification," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2237–2241, sep 2018.

[20] X. Huang, A. Acero, and H. Hsiao-Wuen, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*. Philadelphia, PA: Prentice Hall, apr 2001.

[21] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, jan 2010.

[22] 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," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, nov 2012.

[23] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, no. 5, pp. 1065–1068, 2007.

[24] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[25] M. Slaney and R. F. Lyon, "A perceptual pitch detector," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, pp. 357–360, 1990.

[26] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, dec 2013, pp. 273–278.

[27] E. M. Zion Golumbic, D. Poeppel, and C. E. Schroeder, "Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective,"

Brain and Language, vol. 122, no. 3, pp. 151–161, sep 2012.

[28] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," Meta Research, 2021.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

[30] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," 2013.

[31] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very Deep Convolutional Neural Networks for Raw Waveforms," *International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, pp. 3–7, oct 2016.

[32] I. Tokuda, "The Source–Filter Theory of Speech," in *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, nov 2021, pp. 1–15.

[33] I. R. Titze, *Principles of Voice Production*. Prentice Hall, 1994.

[34] B. Bogert, H. M.J., and T. J.W., "The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking," in *Proc. Symposium Time Series Analysis, 1963*. New York: Wiley, 1963, pp. 209–243.

[35] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley Publishing Company, 1987.

[36] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, aug 1980.

[37] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, 2015.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[41] J. S. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," *Advances in Neural Information Processing Systems*, vol. 2, no. Ml, pp. 211–217, 1990.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[43] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing: principles, algorithms, and applications*, 4th ed. Prentice Hall, 2006.

[44] K. Paliwal and K. Wojcicki, "Effect of Analysis Window Duration on Speech Intelligibility," *IEEE Signal Processing Letters*, vol. 15, pp. 785–788, feb 2008.

[45] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, oct 2010.

[46] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[47] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *Eurasip Journal on Advances in Signal Processing*, no. 1, 2015.

[48] J. P. Egan, *Signal detection theory and ROC-analysis*. Academic press, 1975.

[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[50] D. Freeman, C. Southcott, G. Cosier, D. Sereno, A. van der Krogt, A. Gilloire, and H. Braun, "Voice control of the pan-European digital mobile radio system," in *IEEE Global Telecommunications Conference, 1989, and Exhibition. 'Communications Technology for the 1990s and Beyond*. IEEE, 1989, pp. 1070–1074.

[51] F. Beritelli, S. Casale, and A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1818–1829, 1998.

[52] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, no. 3, pp. 245–254, 1995.

[53] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Improved performance measures for voice activity detection," *Proceedings of 11th ITG Symposium on Speech Communication*, pp. 24–27, sep 2014.

[54] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[55] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication*, vol. 20, no. 1-2, pp. 13–22, nov 1996.

APPENDIX A
NETWORK TRAINING AND HYPERPARAMETER VALIDATION

The LSTM networks are trained on 8, 16 or 24 MFCCs from the 60% of the WSA set reserved for training. In an effort to increase robustness towards noise, some networks have 20% of the WSA training data replaced with speech-in-noise mixtures varying between -20 and 20 dB SNR. These are generated using TIMIT[29] training sentences and DEMAND[30] cafeteria noise, and are much more speech-sparse than the WSA recordings (15% versus 31% speech per file). Training data size is kept constant across all networks, and equates to around 45 hours of recordings. Training is done on segments of MFCCs whose length is part of determining the temporal context available to the network during training. Sequences are grouped together in *mini-batches* consisting of 32 sequences at a time. Each mini-batch is normalized individually. Training is concluded when the network has completed 10 full passes of the training set, shuffled after each pass.

To get a better understanding of the effect of certain architecture and training choices, a selection of networks are trained and validated against each other. The parameters to validate are: batch-normalization, dropout, network depth, length input sequences, number of MFCCs and the benefit of using biLSTM over LSTM. The validation process consists of evaluating hit-rate minus false alarm and accuracy on speech files that are disjoint from data used in training and testing. A subset of the WSA data equating to 10% of the full set has been reserved for validation. The networks will be validated both on this data alone and on a 80:20 mixture of WSA validation data and speech-in-noise (using unseen speech and restaurant noise). Fig. 8 shows the result of validating network parameters for the networks summarized in Table 2.

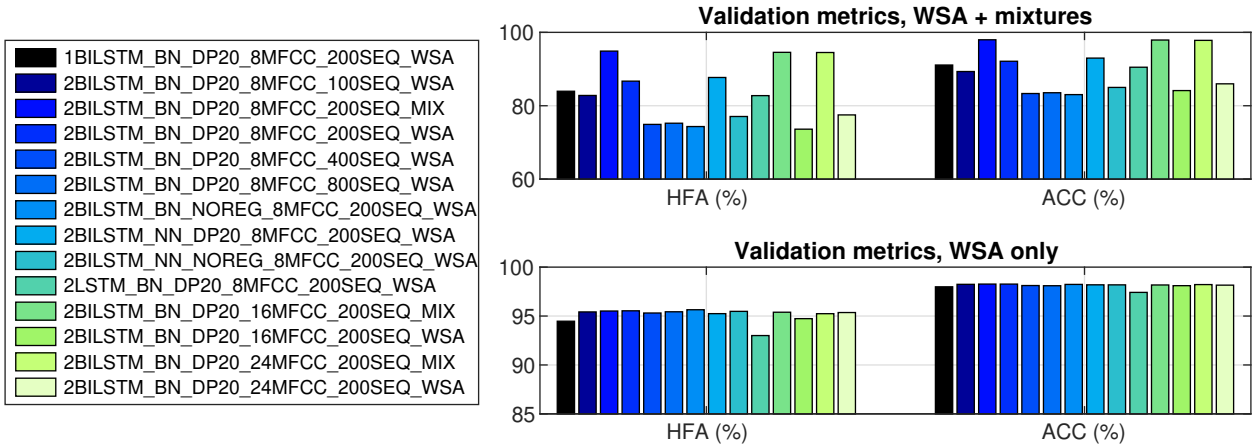

Figure 8. VAD network validation HFA and ACC for all network parameter combinations. Top: Validation results on WSA with 20% of files replaced with TIMIT+DEMAND speech-in-noise mixtures. Bottom: Validation on WSA data purely. The networks trained on mixed data perform notably better on a mixed validation set without losing performance on the target data. Further, there is a net gain to using biLSTM in favor of LSTM, as well as using a training sequence length of 200 frames. Both regularization and dropout are also shown to be beneficial. The number of MFCCs does not seem to influence performance when all else is kept equal.

Table 2
Comprehensive list of networks with selected hyperparameters for validation. Results can be seen in Fig. 8

Network name	Batchnorm	Regularization	# of MFCCs	Seq. length	Training data
1BILSTM_BN_DP20_8MFCC_200SEQ_WSA	yes	yes	8	200	WSA
2BILSTM_BN_DP20_8MFCC_100SEQ_WSA	yes	yes	8	100	WSA
2BILSTM_BN_DP20_8MFCC_200SEQ_MIX	yes	yes	8	200	WSA+MIX
2BILSTM_BN_DP20_8MFCC_200SEQ_WSA	yes	yes	8	200	WSA
2BILSTM_BN_DP20_8MFCC_400SEQ_WSA	yes	yes	8	400	WSA
2BILSTM_BN_DP20_8MFCC_800SEQ_WSA	yes	yes	8	800	WSA
2BILSTM_BN_NOREG_8MFCC_200SEQ_WSA	yes	no	8	200	WSA
2BILSTM_NN_DP20_8MFCC_200SEQ_WSA	no	yes	8	200	WSA
2BILSTM_NN_NOREG_8MFCC_200SEQ_WSA	no	no	8	200	WSA
2LSTM_BN_DP20_8MFCC_200SEQ_WSA	yes	yes	8	200	WSA
2BILSTM_BN_DP20_16MFCC_200SEQ_MIX	yes	yes	16	200	WSA+MIX
2BILSTM_BN_DP20_16MFCC_200SEQ_WSA	yes	yes	16	200	WSA
2BILSTM_BN_DP20_24MFCC_200SEQ_MIX	yes	yes	24	200	WSA+MIX
2BILSTM_BN_DP20_24MFCC_200SEQ_WSA	yes	yes	24	200	WSA