IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

The observation likelihood of silence: analysis and prospects for VAD applications

Igor Odriozola[1], *Inma Hernaez*[1], *Eva Navas*[1], *Luis Serrano*[1], *Jon Sanchez*[1]

[1]AHOLAB, University of the Basque Country UPV/EHU
{igor,inma,eva,lserrano,ion}@aholab.ehu.eus

Abstract

This paper shows a research on the behaviour of the observation likelihoods generated by the central state of a *silence* HMM (Hidden Markov Model) trained for Automatic Speech Recognition (ASR) using cepstral mean and variance normalization (CMVN). We have seen that observation likelihood shows a stable behaviour under different recording conditions, and this characteristic can be used to discriminate between *speech* and *silence* frames. We present several experiments which prove that the mere use of a decision threshold produces robust results for very different recording channels and noise conditions. The results have also been compared with those obtained by two standard VAD systems, showing promising prospects. All in all, observation likelihood scores could be useful as the basis for the development of future VAD systems, with further research and analysis to refine the results.

Index Terms: VAD, observation likelihood, cepstral normalization

1. Introduction

Voice activity detection (VAD) is an important issue in Automatic Speech Recognition (ASR) or ASR-based systems. It allows the systems to reduce the computation cost and, as a consequence, the response time of the decoding process, by only passing speech frames [1]. If the access to the system is intended to be universal, the VAD has to cope with different noise levels, with no —or little— loss in accuracy. Indeed, the greatest challenge for the current ASR systems is to cope with background noise in the input speech signal [2].

A large number of speech features and combinations have been proposed for VAD [3]. Gaussian Mixture Models (GMM) and Hidden Markov Models (HMMs) have been tested in this context [4][5]. Recently, the use of classifiers has been very common: decision trees (DT) [6], Support Vector Machines (SVM) [7] and hybrid SVM/HMM architectures [8]. More recently, neural networks (NN) have appeared in the literature outperforming the previous designs [9][10][11]. However, these approaches are complex and do not work in real time.

Little research has been done using cepstral normalization for VAD proposals, although it proved to be rather discriminative already in [12]. Here, we introduce some research on the use of observation likelihoods for VAD, applying Cepstral Mean and Variance Normalization (CMVN). We analyse the behaviour of the observation likelihoods generated by the GMM in the central state of the *silence* HMMs trained for ASR. Results show that it is a promising basis for future prospects.

The next section is a study of different aspects of the observation likelihood scores. Section 3 describes the databases and metrics used for the experiments. Then, VAD some experiments are shown in section 4. Finally, some conclusions and future prospects are explained in section 5.

2. The observation likelihood

In speech recognition, audio segments corresponding to the same recognition unit (word, phone, triphone etc., even *silence* or non-speech) are gathered and processed, in order to extract acoustic parameters from them —typically Mel-frequency cepstral coefficients (MFCC)— and train a different acoustic model for each unit. A very popular acoustic model is the HMM, since it not only models the likelihood of a new observation vector, but also the sequentiality of the observations.

Usually, observation likelihoods are generated by the GMM belonging to each HMM state j. For an observation vector o_t, the observation likelihood b_j of a GMM is calculated as shown in equation 1.

$$b_j\left(o_t\right) = \sum_{m=1}^{M} c_{jm} N\left(o_t; \mu_{jm}, \Sigma_{jm}\right) \quad (1)$$

where M is the number of mixture components, c_{jm} is the weight of the m^{th} component and $N(\cdot;\mu;\Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ.

In this work, the observation likelihoods have been obtained from the *silence* HMM trained using the *Basque Speecon-like* database [13], specifically the *close-talk* channel.

2.1. The acoustic model for silence

The HMM topology chosen for *silence* frames has three states, left-to-right, allowing the right-end state to connect back with the left-end state. It was trained with 13 MFCCs and 13 first and 13 second order derivatives as acoustic parameters, and 32-mixtures GMMs. The frame length is 25 ms with a shift of 10 ms.

CMVN was applied to MFCCs, computing global means and variances from each recording session. For N cepstral vectors $y = \{y_1, y_2, ..., y_N\}$, their mean μ_N and variance σ_N^2 vectors are calculated as defined in equations 2 and 3, respectively.

$$\mu_N\left(i\right) = \frac{1}{N}\sum_{n=1}^{N} y_n\left(i\right) \quad (2)$$

$$\sigma_N^2\left(i\right) = \frac{1}{N}\sum_{n=1}^{N} \left(y_n\left(i\right) - \mu_N\left(i\right)\right)^2 \quad (3)$$

where i is the i^{th} component of the vector.

The cepstral features are then normalized using the calculated mean and variance vectors, as given in equation 4. Thus, each normalized feature has zero mean and unit variance.

$$\hat{y}_n\left(i\right) = \frac{y_n\left(i\right) - \mu_N\left(i\right)}{\sigma_N\left(i\right)} \quad (4)$$

10.21437/IberSPEECH.2018-11

2.2. The impact of CMVN

The use of CMVN has a significant impact on the curves that observation likelihoods form. When testing a sample signal and computing frame by frame the observation likelihoods at each state of the *silence* HMM, very different curves are obtained depending on weather CMVN is applied or not. Figure 1 illustrates this difference. The middle and bottom diagrams show the curves formed by the observation log-likelihoods generated by each HMM state s_0, s_1 and s_2, without and with normalization respectively, through a utterance composed of four words. In this case, the normalization has been performed using the means and variances computed from the file.

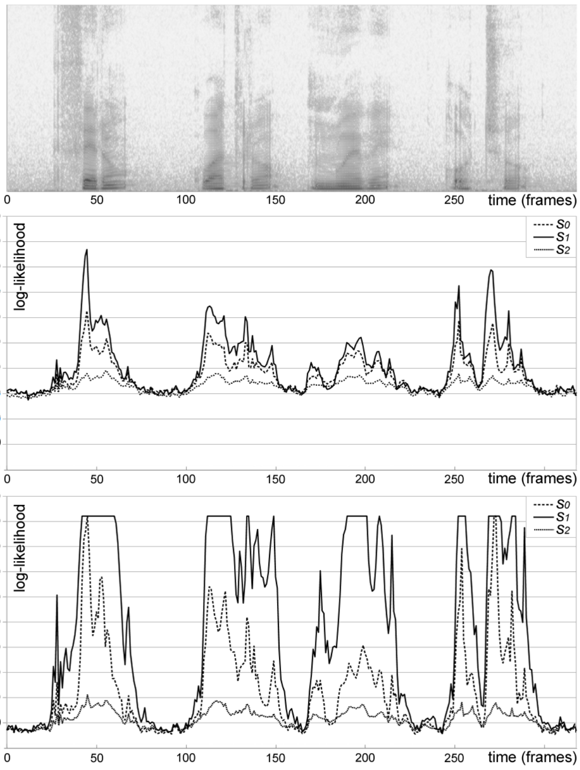

Figure 1: *Spectrogram (top) and observation log-likelihoods along time (frames) generated by the left state (s_0), central state (s_1) and right state (s_2) of the silence HMM without CMVN (middle) and applying CMVN (bottom).*

The curves in the bottom diagram (with CMVN), compared with the ones in the middle diagram (without CMVN), look more abrupt. This fact can be used to better discern between speech and non-speech.

2.3. The central state of the *silence* HMM

In any three-state HMM, the central state is a priori the most stable state of the model, since the left and right states have to cope with transitions between models. It makes sense that the same will happen to the *silence* HMM, where left and right states have to model transitions between silence and speech.

Looking back at Figure 1, we can see that, indeed, the curves generated by the central state (s_1) are, in both cases (with and without cepstral normalization), much more discriminative than the curves corresponding to the states at the ends, which are more irregular.

2.4. Robustness against different SNR values

Another interest point to focus on in a VAD is its robustness for different recording conditions. As an example, we have chosen four signals from the *Spanish SpeeCon* database [14] to illustrate the impact of the recording distance on the observation likelihood curves. These four signals correspond to the same utterance, but were recorded by means of four different microphones: a headset (channel C_0), a lavalier (channel C_1), a medium-distance cardioid microphone (0.5-1 meter, channel C_2) and a far-distance omnidirectional microphone (channel C_3). Each of these channels represents a different SNR, C_0 being the cleanest (around $20dB$) and C_3 the noisiest ($0dB$).

Figure 2 shows the observation log-likelihoods generated by the central state of the *silence* HMM trained with the *Basque Speecon-like* database. The utterance is the same as the one in Figure 1 (note that the signal in Figure 1 corresponds to the C_1 signal in Figure 2). The darkest curve corresponds to the C_0 channel and the lightest one to the C_3 channel.

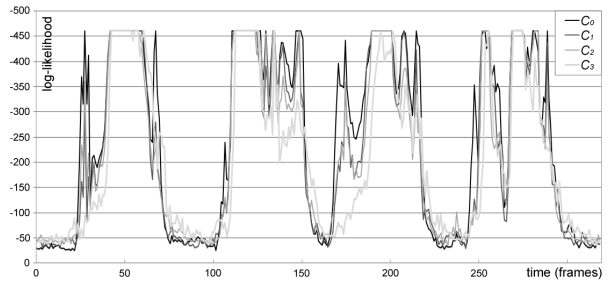

Figure 2: *Observation log-likelihoods along time obtained at the central state (s_1) of the silence HMM when processing different channels (C_0, C_1, C_2, C_3).*

The curves show that, as expected, a degradation occurs when the signals recorded at farther distances are processed, but even so the curves remain rather discriminative. For C_3 signals, the most adverse effect occurs at the initial and ending phones, where, depending on the phone, likelihoods can be very similar to those of the noisy silence. This happens mostly when the initial phone is a noisy phone. However, the curves show a good behaviour for C_1 and C_2, with likelihood profiles very similar to those obtained for C_0 signals.

3. Data preparation

To assess the stability of the observation likelihood curves generated by the central state of the *silence* HMM, a VAD accuracy experiment has been carried out, setting different thresholds to label frames as *speech* or *silence*.

3.1. The databases

Two databases have been chosen for the experiments: first, the *Noisy TIMIT spech database* [15], to analyse weather a threshold could be set for different SNR conditions. The second database is the *ECESS* subset of the *Spanish Speecon database* [16], which has been used to test the validity of that threshold.

1. *Noisy TIMIT spech database*: it contains approximately 322 hours of speech from the TIMIT database [17] modified with different additive noise levels. However, we have chosen only babble and white noises, as the most natural ones. Noise levels vary in 5 dB steps and ranging from 50 to 5 dB. The database contains 630 different

speakers, with 10 utterances per speaker: 6300 files for each noise level. The total speech content in the database is 86.57 % (not well balanced), and the label files are the ones belonging to the classic TIMIT database. All audio files are presented as single channel 16kHz 16-flac, but have been converted to 16-bit PCM.

2. *ECESS* subset of the *Spanish Speecon database*: it was used in the ECESS evaluation campaign of voice activity and voicing detection in 2008. It includes 1020 utterances recorded in different environments (office, entertainment, car and public place) distributed among the C_0, C_1, C_2 and C_3 subsets (total number of files: 4080). There are 60 different speakers each of which utters 17 sentences. The total speech content in the database is 55.77 % (well balanced), and it contains reference speech and silence labels specifically designed to assess different VAD algorithms. The signals in the database were recorded at 16 kHz and 16 bit per sample.

Each file's features have been normalized *off-line*, with the means and variances calculated from the file itself. The *on-line* performance has been left for future research.

3.2. Error metrics

The VAD accuracy experiment consists in evaluating the ability of the system to discriminate between speech and silence segments at different SNR levels, in terms of silence error-rate (ER_0) and speech error-rate (ER_1). These two rates are computed as the fractions of the silence frames and speech frames that are incorrectly classified ($N_{0,1}$ and $N_{1,0}$, respectively) among the number of real silence frames and speech frames in the whole database (N_0^{ref} and N_1^{ref}, respectively), as shown in equation 5. In addition, the TER (total error rate) has also been computed as the average of the ER_0 and ER_1 (equation 6).

$$ER_0 = \frac{N_{0,1}}{N_0^{ref}} \times 100; ER_1 = \frac{N_{1,0}}{N_1^{ref}} \times 100 \quad (5)$$

$$TER = \frac{ER_0 + ER_1}{2} \quad (6)$$

A minimum duration of 15 frames both for speech and silence segments was set. This value was empirically chosen after some preliminary experiments.

4. VAD experiments

Initially, we have analysed whether a threshold can be set for VAD purposes, considering the various SNR values. Then, we have tested that threshold in a separate database, and, in addition, a validity test has been carried out comparing the results with those obtained with three standard VAD algorithms.

4.1. Analysis of the decision threshold

Different thresholds have been considered to label frames as *speech* or *silence*. Results are shown in Figure 3, both for *babble* noise (left) and *white* noise (right).

For the cleanest signals ($SNR = 50dB$), the equal error rate (EER) points of ER_0 and ER_1 curves are located near -200. However, as the SNR gets lower, the EER points move towards higher values. In the case of *white* noise, this shift reaches the -120 value for 5 dB.

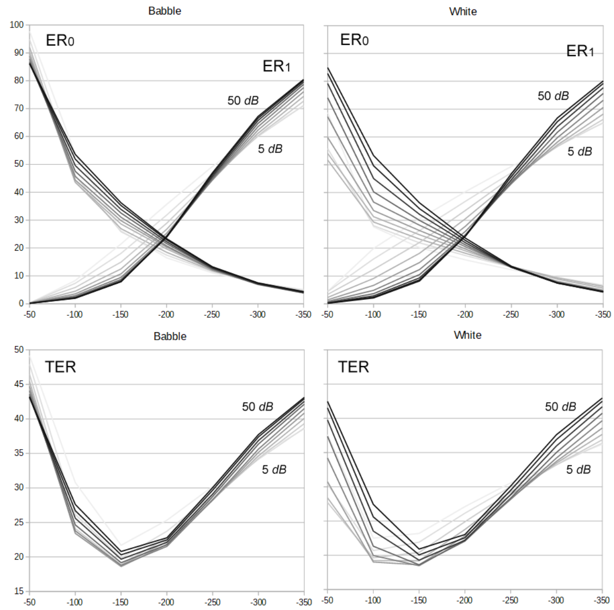

Figure 3: *ER_0 and ER_1 (top) and TER (bottom) for different decision threshold values when testing the signals of SNR 50 to 5 dB in the babble noise subset (left) and the white noise subset (right) of the Noisy TIMIT database.*

Regarding the error rates, the minimum TERs are obtained at $Th = -150$, except for 5, 10 and 15 dB in *white* noise subset, which occur at -100. Thus, we can consider the point of $Th = -150$ as the most valid threshold. Some ER_0 and ER_1 values obtained for $Th = -150$ are shown in Table 1.

Table 1: *TER, ER_0 and ER_1 for $Th = -150$ on the signals of SNR 50, 35, 20 and 5 dB in the babble noise (left) and white noise (right) subsets of the Noisy TIMIT database.*

	Babble			White		
	ER_0	ER_1	TER	ER_0	ER_1	TER
50dB	34.89	6.71	20.80	34.88	6.95	20.92
35dB	30.87	7.48	19.18	28.05	9.18	18.62
20dB	26.35	11.25	18.80	21.53	16.89	19.21
5dB	22.60	20.78	21.70	15.49	30.90	23.20

For $Th = -150$, the minimum ER_1 is 6.71, at 50 dB. As expected, the ER_1 increases as the SNR decreases. However, notice that the TER does not present the minimum at 50 dB, neither in the *babble* noise subset nor in the *white* noise subset, as might be expected.

4.2. Testing

The threshold calculated in the previous section has been applied to the files of *ECESS* subset of the *Spanish Speecon database*. 4080 files have been tested (1020 in each C_i subset). Results are shown in Table 2.

The results obtained for the *ECESS* subset using the threshold calculated from the *Noisy TIMIT* are very good. Compared with the best result obtained for the *Noisy TIMIT* (see 50 dB row in Table 1), much lower ER_0 and ER_1 have been obtained. The error rates, as expected, increase as SNR decreases, al-

Table 2: TER, ER_0 and ER_1 with $Th = -150$ *on the signals of channels* C_0, C_1, C_2 *and* C_3 *in the Spanish Speecon database.*

	ER_0	ER_1	TER
C_0	6.21	2.74	4.48
C_1	4.22	6.13	5.18
C_2	7.10	6.00	6.55
C_3	9.46	6.45	7.96

though the best silence error rate is obtained for the C_1 channel.

Additionally, a tuning has been performed for ER_1 reduction. Indeed, for speech processing, it is important to reduce the ER_1 as much as possible, so that the minimum number of speech frames are lost for the next stage. For that purpose, we have sought to reduce the impact of non-speech to speech boundaries, setting an additional margin of 5 and 10 frames around the speech segments. Results are shown in Table 3.

Table 3: TER, ER_0 *and* ER_1 *for 5 and 10 frames long speech-segment margins, with* $Th = -150$ *for the signals of channels* C_0, C_1, C_2 *and* C_3 *in the ECESS subset of the Spanish Speecon database.*

	5 frames			10 frames		
	ER_0	ER_1	TER	ER_0	ER_1	TER
C_0	10.84	1.29	6.07	15.68	0.79	8.24
C_1	7.94	3.47	5.71	12.42	2.30	7.36
C_2	10.91	3.50	7.21	15.39	2.47	8.93
C_3	13.29	3.95	8.62	17.59	2.89	10.24

The table shows that ER_1 reduces and ER_0 increases. TER increases as well, because ER_0 increases faster than ER_1 reduces. All in all, the use of a margin around speech segments allows decreasing significantly ER_1, with a not very significant resulting TER degradation.

4.3. Comparison with other systems

In order to validate the previous results, our results have been compared with the outcomes of three popular standard VAD algorithms carried out in a previous work [18]. These systems are standard defined by ITU (International Telecommunication Union) and ETSI (European Telecommunications Standards Institute):

1. The VAD algorithm of the ITU G.729 system [19].
2. The AFE-FD (frame-dropping mechanism) algorithm implemented in ETSI AFE-DSR (Advanced Front-End for Distributed Speech Recognition) [20].
3. The AFE-NR (noise reduction system) algorithm implemented in ETSI AFE-DSR [20].

Table 4 shows the results obtained for the three VAD systems along with the proposed method (using $Th = -150$ and a margin of 10 frames), over the same dataset (4080 files from the ECCESS subset). Regarding ER_1, the AFE-FD gets better results, and also the AFE-NR for C_0 and C_1. However both systems show the disadvantage of getting very high ER_0 for all the channels (the lowest value is 38.10 %). This means that many silence frames will be sent to the recognizer. The ER_0 in our results are between 12.42 and 17.59 %.

Table 4: *Comparison of different VAD algorithm results at four* SNR *levels*

(a) Silence error rates (ER_0)

	G.729	AFE-FD	AFE-NR	Prop.
C_0	56.06	63.88	58.23	15.68
C_1	70.23	54.75	55.96	12.42
C_2	59.54	52.10	38.10	15.39
C_3	70.49	50.10	47.65	17.59

(b) Speech error rates (ER_1)

	G.729	AFE-FD	AFE-NR	Prop.
C_0	3.63	0.03	0.62	0.79
C_1	9.28	0.23	1.98	2.30
C_2	18.19	0.48	4.83	2.47
C_3	17.22	1.41	8.30	2.89

5. Conclusions

In this paper, we have assessed the usefulness of the observation likelihood generated by the central state GMM of a *silence* HMM trained using CMVN, as a possible basis on which to build a VAD system. We have seen that a good classification between *speech* and *silence* can be performed, just by setting a threshold in the curves that observation likelihoods form.

The *silence* HMM has been trained using the *close-talk* channel from the *Basque Speecon-like* database. Then, a threshold analysis has been carried out, processing the *babble* and *white* noise files of the *Noisy TIMIT* database. As a conclusion, we have noticed that the minimums error rates occur at the same likelihood point in 17 SNR values out of a total of 20. This point is the one we have chosen as the threshold.

This threshold has been tested with a separate database: the *ECESS* subset of the *Spanish Speecon* database. The results obtained for this database are even better than those obtained for the *Noisy TIMIT*, which leads us to think that the *silence* observation likelihood behaves similarly on different channels.

Additionally, the results of the test have been compared with three different standard VAD systems. Although the best speech error rates have not been achieved with the use of the decision threshold, we have got the best silence error rates. Our results are quite competitive; actually, the best total classification rates have been obtained.

As a final conclusion, competitive results are obtained just by setting a decision threshold to the *silence* observation likelihood curves. This fact has been applied in [21], where a method called Multi-Normalization Scoring (MNS) is used to explode the discriminative potential of the observation likelihood scores. Robust on-line results are shown in that paper, where the scores obtained with MNS are classified with a Multi-Layer Perceptron (MLP). This issue and others related to the selection of the optimal threshold are being investigated currently in our laboratory.

6. Acknowledgements

This work has been partially supported by the EU (FEDER) under grant TEC2015-67163-C2-1-R (RESTORE) (MINECO/FEDER, UE) and by the Basque Government under grant KK-2017/00043 (BerbaOla).

7. References

[1] M. K. Mustafa, T. Allen, and K. Appiah, *Research and Development in Intelligent Systems XXXI*. Springer International Publishing, 2014, ch. A Review of Voice Activity Detection Techniques for On-Device Isolated Digit Recognition on Mobile Devices, pp. 317–329.

[2] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*, 1st ed. Wiley Publishing, 2012.

[3] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, 2000.

[4] J. Tatarinov and P. Pollák, "Hmm and ehmm based voice activity detectors and design of testing platform for vad classification," *Digital Technologies*, vol. 1, pp. 1–4, 2008.

[5] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," vol. 6, no. 1, pp. 54–63, 2012.

[6] Ó. Varela, R. S. Segundo, and L. A. Hernández, "Combining pulse-based features for rejecting far-field speech in a HMM-based Voice Activity Detector," vol. 37, no. 4, pp. 589–600, 2011.

[7] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *ICSP 2002 – 6th International Conference on Signal Processing Proceedings, August 26-30, Beijing, China, Proceedings*. IEEE, 2002, p. 464467.

[8] Y. W. Tan, W. J. Liu, W. Jiang, and H. Zheng, "Hybrid svm/hmm architectures for statistical model-based voice activity detection," 2014, pp. 2875–2878.

[9] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," 2013, pp. 7378–7382.

[10] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," 2014, pp. 2519–2523.

[11] Y. Obuchi, "Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression," in *ICASSP*, 2016, pp. 5715–5719.

[12] M. Westphal, "The use of cepstral means in conversational speech recognition," in *EUROSPEECH 1997 – 5th European Conference on Speech Communication and Technology, September 22-25, Rhodes, Greece, Proceedings*. ISCA, 1997, pp. 1143–1146.

[13] I. Odriozola, I. Hernaez, M. I. Torres, L. J. Rodriguez-Fuentes, M. Penagarikano, and E. Navas, "Basque speecon-like and Basque speechdat MDB-600: speech databases for the development of ASR technology for Basque," in *LREC 2014, Ninth International Conference on Language Resources and Evaluation, May 26-31, Reykjavik, Iceland, Proceedings*, 2014, pp. 2658–2665.

[14] D. Iskra, B. Grosskopf, K. Marasek, H. van den, F. Diehl, and A. Kiessling, "Speecon speech databases for consumer devices: Database specification and validation," in *LREC 2002, Third International Conference on Language Resources and Evaluation, May 27-31, Las Palmas, Spain, Proceedings*, 2002, pp. 329–333.

[15] A. Abdulaziz and V. Kepuska, "Noisy timit speech (ldc2017s04)," 3 2017. [Online]. Available: http://hdl.handle.net/11272/UFA9N

[16] B. Kotnik, P. Sendorek, S. Astrov, T. Koç, T. Çiloglu, L. D. Fernández, E. R. Banga, H. Höge, and Z. Kacic, "Evaluation of voice activity and voicing detection," in *INTERSPEECH 2008 – 8th Annual Conference of the International Speech Communication Association, September 22-26, Brisbane, Australia, Proceedings*, 2008, pp. 1642–1645.

[17] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[18] I. Luengo, E. Navas, I. Odriozola, I. Saratxaga, I. Hernaez, I. Sainz, and D. Erro, "Modified LTSE-VAD algorithm for applications requiring reduced silence frame misclassification." in *LREC 2010, Seventh International Conference on Language Resources and Evaluation, May 17-23, Valletta, Malta, Proceedings*, 2010, pp. 1539–1544.

[19] P. Setiawan, S. Schandl, H. Taddei, H. Wan, J. Dai, L. B. Zhang, D. Zhang, J. Zhang, and E. Shlomot, "On the itu-t g.729.1 silence compression scheme." in *EUSIPCO 2008 – 16th European Signal Processing Conference, August 25-28, Lausanne, Switzerland, Proceedings*, 2008, pp. 1–5.

[20] E. Standards, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *ETSI Standards, European Telecommunications Standards Institute*, vol. ES 201 108 Recommendation, 2002.

[21] I. Odriozola, I. Hernaez, and E. Navas, "An on-line VAD based on Multi-Normalisation Scoring (MNS) of observation likelihoods," *Expert Systems with Applications (ESwA)*, vol. 110, pp. 52–61, 2018.