Comparison of Forced-Alignment Speech Recognition and Humans for Generating Reference VAD

Ivan Kraljevski[1], *Zheng-Hua Tan*[2], *Maria Paola Bissiri*[3]

[1]voice INTER connect GmbH, Dresden, Germany
[2]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
[3]CASL Research Centre, Queen Margaret University, Edinburgh, United Kingdom
ivan.kraljevski@voiceinterconnect.de, zt@es.aau.dk, MBissiri@qmu.ac.uk

Abstract

This present paper aims to answer the question whether forced-alignment speech recognition can be used as an alternative to humans in generating reference Voice Activity Detection (VAD) transcriptions. An investigation of the level of agreement between automatic/manual VAD transcriptions and the reference ones produced by a human expert was carried out. Thereafter, statistical analysis was employed on the automatically produced and the collected manual transcriptions. Experimental results confirmed that forced-alignment speech recognition can provide accurate and consistent VAD labels.

Index Terms: voice activity detection, speech recognition, speech segmentation

1. Introduction

Voice activity detection (VAD) attempts to distinguish the presence or absence of human speech in an acoustic signal. VAD is used as a front-end component in many speech-enabled systems, like in robust speech recognition, speech coding and compression systems for low-bandwidth transmission. Detected non-speech segments can be subsequently discarded to improve the overall performance of such systems, saving on computation and on network bandwidth [1].

In general, VAD implements feature calculation and classification of an acoustic signal segment as speech or non-speech. Standard VAD methods are based on energy thresholds (non-adaptive and adaptive), waveform and spectrum analysis (pitch and harmonic detection, periodicity measures, zero crossing rate, spectral entropy, etc.) [1] or on statistical models [2],[3]. Another type of VAD is supervised VAD, which requires a large amount of speech data labeled by humans, i.e. data along with VAD transcriptions [1].

To evaluate a VAD system, its output produced on a test corpus is compared with a reference VAD, which is created manually by humans. Producing reference VAD is costly and time consuming and in some cases, not possible at all, such as for very large speech databases. Moreover, labeling carried out on the same speech corpus by different persons (including experts) can lead to significant differences, inconsistent and erroneous transcriptions.

Reference VAD can also be generated by using energy-based or statistical VAD on a clean speech corpus. In this case, the evaluation of VAD methods is performed on noisy versions of the speech corpus. These approaches certainly introduce bias towards one class of VAD methods while evaluating the methods.

The quality of human VAD labelling is frequently neglected and the inconsistency of human annotations makes it difficult to reliably interpret experimental results. On the other hand, forced-alignment automatic speech recognition (ASR) is more consistent although dependent on the acoustic model as well. ASR systems are extensively used for the initial segmentation of speech. A HMM based phonetic recognizer is commonly employed for phoneme segmentation and for estimating the phoneme boundaries by means of Viterbi forced-alignment [5],[6].

While there are numerous studies dealing with the accuracy of automatic versus manual phoneme segmentation, e.g. [7],[8], to our best knowledge the agreement between forced-alignment ASR and manually provided VAD transcriptions has not been investigated in the literature.

Based on some preliminary investigations and experiments as well as on the results of various studies on phoneme segmentation, e.g. [4],[5], we hypothesize that Viterbi forced-alignment for transcribing VAD can be as precise as a human expert and better than most non-expert annotators, at the same time providing consistent VAD labels. Automatic VAD transcription is commonly used without any systematic study [9],[10], and the present work should provide a foundation for employing it.

Specifically, we investigate whether forced-alignment speech recognition can be an alternative to human annotations in generating reference VAD. The level of agreement of automatic as well as of manual VAD transcriptions with the reference ones (generated by a human expert) was investigated. A set of sentences was prepared and experiment participants were asked to perform VAD annotations. Statistical analysis was carried out on the automatically produced and the collected manual transcriptions.

The paper is organized as follows: Section 2 describes the creation of the automatic VAD transcriptions, the speech databases used for acoustic modeling, the feature extraction and the speech recognition engine that was used. Section 3 presents the methodology of the human labeling experiment, Section 4 the results of the statistical analysis, and Section 5 the conclusions.

2. Automatic VAD Transcription

2.1. Speech and Language Databases

For ASR acoustic modeling, the database we used comprises mixed speech corpora in German: Phonedat I read speech corpus [11] and Verbmobile I spontaneous speech corpus [12]. The total duration of the speech database is approx. 53 hours and 15 minutes.

The phoneme transcription of the words from the original database was generated by means of an automatic grapheme-to-phoneme (G2P) procedure and included in the training dictionary in order to ensure consistent pronunciations for acoustic model training. The G2P model that we used was trained on a lexicon derived from the WebCelex database [13].

The speech data used in the VAD transcription experiments consist of 34 sentences recorded by a native German speaker in a studio environment. The data were selected from a domain-specific corpus recorded for investigating pathological speech and pronunciation errors by speech-impaired patients [14].

2.2. ASR system

For the forced-alignment segmentation and VAD labeling, the Sphinx/pocketsphinx [15] framework and its Gstreamer [16] realization were employed without using the VAD functionality. Acoustic modeling was conducted using Sphinx training tools and customized procedures for model training and testing were established.

The basic requirement for forced-alignment ASR-based VAD transcription is that the corresponding phoneme sequence has to be known in advance. The recognizer is configured for phoneme recognition, where the pronunciation dictionary that is used consists only of phonemes.

Forced-alignment is performed by means of separate finite-state grammars containing a single state sequence without alternatives. Pauses between words are optional, as is presence/inclusion of the glottal stop, thus providing a more accurate speech/non-speech segmentation.

2.3. Feature extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as standard features with different stream types depending on the acoustic model employed (continuous or semi-continuous).

Some of the parameters we used were: pre-emphasis coefficient of 0.97, Hamming window of 32 ms, frame rate of 10 ms, cepstral mean normalization (CMN) subtracting an average computed over the whole processed utterance, filter-bank with 40 overlapping frequency bands with triangular response for 16 kHz, 13 MFCCs with c0, completed by dynamic and acceleration coefficients for single stream features continuous acoustic models.

2.4. Acoustic models

The standard procedure for acoustic modeling was used with additional modifications:

1) Forced-alignment was employed to properly align the transcriptions to the utterances prior to training, providing better acoustic modeling by excluding the sentences that could not be aligned. Excluding non-aligned sentences resulted in more consistent data used in the training.

2) Linear Discriminative Analysis combined with Maximum Likelihood Linear Transformation as feature-space transformations provided word error rate (WER) reduction (up to 25% relative in some of the tests).

A number of acoustic models were trained with different training configurations: mono-phone, tri-phone, different feature streams, a different number of Gaussian distributions (4-32) and of senones (1000 and 4000), i.e. sets of Gaussian mixtures.

3. Experimental set-up

3.1. VAD Transcription Experiments

In general, VAD annotators label speech data using the implicit rules they believe and these rules are rarely documented. Because of that, the instructions that the participants received for this labeling experiment were kept simple, in order not to introduce too much bias and to reflect real world annotation practices.

In order to collect VAD transcriptions from human annotators with different experiences in speech technologies, a simple labeling protocol was established. Wavesurfer [17] was chosen as the labeling tool, because of its simple but quite powerful user interface, also suitable for less experienced participants. The recordings and the tool were provided in a package containing a user guide to ensure correct procedure for transcriptions. The non-speech segments were labeled with "0" and the speech segments with "1". The participants could visualize the spectrogram and play the audio of the speech.

Normally, in the case of manual labeling by expert phoneticians, labeling criteria are defined beforehand, depending on the purpose of the transcriptions. In the present case, no specific guidelines were provided. To some extent this simulates the reality in creating VAD transcriptions where it is common to simply state that speech data is manually labeled, without mentioning any specific guidelines for the labeling process [3],[18].

The labelers with non-German language background reported no difficulties in carrying out the task on German speech. They reported using audio-visual clues to label speech and non-speech segments. A total of 19 subjects were recruited (age range 22 to 40). Four of them were rejected because their transcriptions were erroneous (using wrong symbols) and could not be used in the statistical analysis. Of the fifteen remaining participants, five declared their level of experience in labeling speech as beginner, three as intermediate, five as advanced and two as proficient (expert).

3.2. Reference transcriptions

One of the proficient labelers was chosen as reference, because of her experience in segmenting and transcribing speech as an expert phonetician (the third author). In the reference transcriptions, silence, breath and other non-speech sounds (cough, clicks, etc.) were categorized as non-speech. However, since there were no specific recommendations regarding the minimal duration of the segments and the way to handle pauses, the reference labeler also transcribed short silent stretches as non-speech.

While labeling plosives, the reference labeler considered their dynamic articulation. Plosives start with a closure, during which the airflow is blocked, indicated by a silent phase in the spectrogram. The closure is followed by a burst, at the point at which the airflow is released, whose acoustic energy is visible in the spectrogram. In some cases, if a plosive was located at the beginning of an utterance preceded by silence, it was not possible to know exactly when the plosive closure started. In this case, the reference label included some silence before the burst, corresponding to the plosive closure, so that the whole plosive duration would approximately match the duration of the following sounds. Non-expert human labelers, not aware of plosive structure, might have excluded plosive closures from the speech labels if this was the first speech activity visible in the spectrogram after silence.

Table 1. *Manual labels for non-speech and speech (VAD 0-1) compared to the reference across increasing level of experience (EXP)*

EXP	VAD	HR (%)	ACC (%)	ST (+)	ST (+)%	SD (+)	ST (-)	ST (-)%	SD (-)
I	0	72.3	31.6	68	13.9	57	58	86.1	42
	1	78.4	20.1	29	34.2	48	62	65.8	32
II	0	90.1	65.4	25	6.7	43	33	93.3	32
	1	93.8	66.8	17	33.3	18	57	66.7	28
III	0	91.7	67.5	54	30.9	47	30	69.1	31
	1	89.7	72.8	18	23.3	19	55	76.7	31
IV	0	94.0	94.5	03	2.3	2	33	97.7	31
	1	93.4	94.5	16	34.9	12	61	65.1	27
ALL	0	85.9	58.6	51	17.3	49	36	82.7	35
	1	87.9	57.5	18	29.1	23	57	70.9	30

The reference labeler labeled onset glottal stops according to the above-mentioned criteria and included a short silent portion before them in the speech labels, indicating the glottal closure. However, glottal stops were considered as non-speech by the ASR (Section 2.2), so this introduced differences between the reference and the ASR labeling.

3.3. Evaluation criteria

In order to evaluate the quality of the VAD transcriptions, the manual and the automatic labels were compared with the reference to measure detection accuracy and the differences in the segment boundaries.

The methodology employed is different compared to the one commonly used in phoneme segmentation, where the accuracy is generally measured in terms of the percentage of the automatic boundaries which are within a given time tolerance from the manually labeled boundaries, as in [4].

Relevant performance indicators for VAD are the speech hit rate (HR_1) and the non-speech hit rate (HR_0) [1]. Hit rate is defined as the ratio of speech frames (respectively non-speech frames) correctly identified as speech frames (respectively non-speech frames):

$$HR_0=N_{0,0} / N_{0ref} , HR_1=N_{1,1} / N_{1ref} \quad (1)$$

where $N_{0,0}$ is the number of speech frames labeled as non-speech, N_{0ref} is the number of frames which are actually non-speech according to the reference transcription. $N_{1,1}$ is the number of frames identified as speech, and N_{1ref} is the number of frames which actually contain speech according to the reference. Values close to 1 for both hit rates are indicators of good speech/non-speech discrimination.

For the evaluation, the following procedure was employed: for each segment of the reference annotations, the closest left boundary in the tested annotations belonging to a segment with the same label was searched for within a 200 ms time tolerance from the reference left boundary. Since the automatic VAD labels are derived from the automatic phonetic segmentation, they were mapped to an appropriate speech/non-speech symbol.

If a label was matched, this was considered as a positive identification, otherwise it was considered to be a negative one. Detection accuracy is defined as the ratio of the number of labels matching the reference and the total number of labels in the reference.

Table 2. *Automatic labels for non-speech and speech (VAD 0-1) compared to the reference across different acoustic models*

	VAD	HR (%)	ACC (%)	ST (+)	ST (+)%	SD (+)	ST (-)	ST (-)%	SD (-)
PHO	0	85.8	73.6	14	11.9	16	65	88.1	35
	1	96.1	73.6	41	31.3	27	38	68.7	20
C104	0	87.0	75.8	15	13.0	10	59	87.0	33
	1	96.2	75.8	30	36.2	22	37	63.8	22
C108	0	86.5	75.8	13	17.4	12	61	82.6	38
	1	96.3	76.9	32	38.6	22	35	61.4	22
C116	0	85.5	72.5	23	13.6	10	67	86.4	39
	1	96.4	73.6	35	38.8	22	36	61.2	22
C132	0	85.3	73.6	18	13.4	10	69	86.6	38
	1	96.4	73.6	38	37.3	22	37	62.7	24
S15C	0	84.9	74.7	7	14.7	6	63	85.3	33
	1	96.4	72.5	30	31.8	22	36	68.2	21
S45C	0	85.2	73.6	8	10.4	7	64	89.6	35
	1	96.2	72.5	33	28.8	22	36	71.2	20
ALL	0	85.7	74.3	14	13.5	11	64	86.5	36
	1	96.3	74.1	34	34.7	23	36	65.3	21

For all labels matching the reference, independently of speech and non-speech segments, the time differences of the left boundaries between the reference and the observed transcriptions were calculated. The time offset of the left boundary was distinguished between positive (the boundary of the tested annotation occurs earlier than the reference annotation boundary) and negative (the boundary of the tested annotation occurs later than the reference annotation boundary).

4. Results

The evaluation results presented in Table 1 were obtained by comparing the reference with the manual VAD transcriptions for non-speech and speech left boundaries (VAD 0-1). They include hit rate, accuracy, mean of positive and negative shifts (ST + and –, boundaries placed earlier and later than the reference respectively) in milliseconds, their percentage and standard deviation.

The participants were divided into groups according to their declared level of experience in labeling speech (I-beginner, II-intermediate, III-advanced and IV-proficient). In total, 476 files were processed, with 2548 manually annotated segments included in the statistical analysis.

From Table 1 it can be seen that the VAD transcriptions provided by the proficient annotator (level IV) match the reference most closely in terms of high and balanced hit rates and detection accuracies for speech and non-speech segments.

Accuracy decreases along with the decrease of experience level. In most cases, labelers placed the left segment boundaries later than the reference (in 70.9% cases for speech and 82.7% for non-speech segments). One reason could be that they did not label the closure portion of plosives as speech (see Section 3.2).

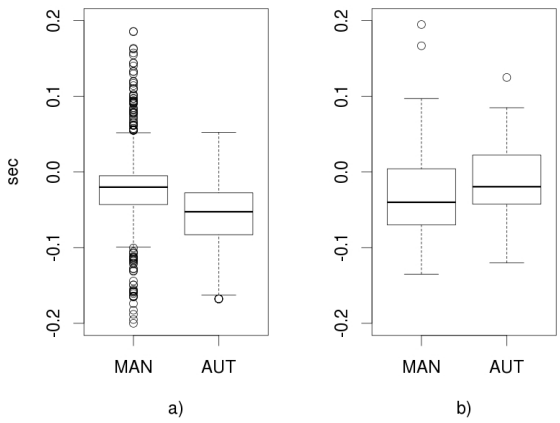

Figure 1: *Distributions of the left boundary differences between the reference labels and the manual and automatic labels for (a) non-speech and (b) speech segments.*

A Shapiro-Wilk test showed that the left boundary differences have non-normal distribution and the Kruskal-Wallis non-parametric test reported significant differences across labeling experience levels ($p<0.001$).

Table 2 presents the result of the comparison of 238 files (that is 1274 automatic annotated labels) with the reference annotation set, for non-speech and speech left boundaries of the annotated intervals (VAD 0-1), across different acoustic models and feature streams, the phoneme model, continuous Gaussian densities of 4, 8, 16, 32 (noted as C104 to C132), and 512 Gaussians semi-continuous model with 1000 (S15C) and 4000 (S45C) senones. The table shows hit rate, accuracy, mean of positive and negative shifts (ST + and –) in milliseconds, their percentage and standard deviation.

The automatic detection accuracies are lower than those by a proficient/expert annotator (level IV), and the hit rates for speech segments are higher than those for non-speech segments. One reason for this difference between the automatic and human annotations might be that the reference transcriptions included several short non-speech segments between words. In the FSG grammars the pauses are defined as optional after each word, and the forced-alignment ASR system could not label such short pauses. Therefore short pauses were missing, introducing more detection errors and unbalanced hit rates. In addition, forced-alignment could produce erroneous results due to incorrect transcriptions, variations in the pronunciation, non-optimal feature extraction (large frame shift etc.) or noisy recordings.

The detection accuracies of the forced-alignment approach are higher than those of human annotators with an advanced level of experience, and far better than those of annotators with an intermediate level of experience and beginners.

From Table 2, it is obvious that automatic transcriptions are more consistent and the left boundary variability is in the same range as in most human transcriptions.

Figure 1 confirms this observation about the consistency of left boundary annotation in automatic and manual as it can be seen that manual VAD transcriptions are characterized by more outliers than the automatic ones for non-speech segments. For the speech segments the automatic labels were closer to the reference. A significant difference was found between manual and automatic boundaries ($p<0.05$, Kruskal-Wallis non-parametric test).

Table 3. *Manual and automatic labels for non-speech and speech (VAD 0-1) compared to the reference after merging segments shorter than 10 and 300 ms with their neighboring segments*

Segment duration threshold of 10 ms									
	VAD	**HR (%)**	**ACC (%)**	**ST (+)**	**ST (+)%**	**SD (+)**	**ST (-)**	**ST (-)%**	**SD (-)**
AUT	0	85.7	74.3	14	13.5	11	64	86.5	36
	1	96.3	74.1	34	34.7	23	36	65.3	21
HUM	0	85.9	58.6	51	17.3	49	36	82.7	35
	1	87.9	57.5	18	29.1	23	57	70.9	30
Segment duration threshold of 300 ms									
AUT	0	89.5	92.7	31	1.3	19	81	98.8	34
	1	98.1	88.0	43	40.8	23	43	59.2	20
HUM	0	90.7	85.1	56	15.2	51	47	84.8	37
	1	89.0	75.1	19	32.1	22	62	67.9	31

As shown also in Tables 1 and 2, forced-alignment speech recognition provided accurate and consistent VAD transcriptions which match the reference better than the transcriptions produced by most human annotators.

In order to avoid errors due to segments that were too short in the transcriptions (see above), a duration threshold was introduced. Manual, automatic and reference labels shorter than a defined threshold (300 ms) were merged with the neighboring segments. It was observed, that the detection accuracy reaches maximum values with a duration threshold around 300 ms, after which it decreases.

Table 3 presents the comparison of human and automatic labels with the reference to the left boundaries of non-speech and speech intervals (VAD 0-1) after merging segments of 10 and 300 milliseconds. The table shows hit rate, accuracy, mean of positive and negative shifts (ST + and –) in milliseconds, their percentage and standard deviation.

In addition, VAD transcriptions by a proficient human annotator (level IV) and the best performing ASR setup (C108) in terms of detection accuracies were compared. No significant differences were found in terms of the left boundaries offsets from the reference (Kruskal-Wallis non-parametric test).

5. Conclusions

The current study compared forced-alignment speech recognition and human generated VAD transcriptions. Statistical analysis was carried out after comparing automatically and manually created transcriptions with the reference ones created by a human expert labeler. It has been shown that forced-alignment can provide transcriptions as good as or even better than most human labelers, matching closely transcriptions made by an expert labeler. We conclude that forced-alignment speech recognition can provide accurate and consistent VAD labels.

Further investigations should be conducted on noisy speech as well as in scenarios where the human annotators are given labeling criteria.

6. Acknowledgments

The authors are thankful to Eleanor Lawson for her comments on an earlier version of this paper.

7. References

[1] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection fundamentals and speech recognition system robustness," *INTECH Open Access Publisher*, 2007

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[3] T. Petsatodis, C. Boukis, F. Talantzis, Z.-H. Tan, and R. Prasad, "Convex combination of multiple statistical models with application to VAD," in *Audio, Speech, and Language Processing, IEEE Transactions on* 19, no. 8 (2011): 2314-2327

[4] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *in Speech and Audio Processing, IEEE Transactions on* 11(6) (2003): 617-625.

[5] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication,* vol. 12, no. 4, pp. 357–370, 1993.

[6] J. W. Kuo, and H. M. Wang, "A minimum boundary error framework for automatic phonetic segmentation," *In Chinese Spoken Language Processing*, *Springer Berlin Heidelberg, 2006,* pp. 399-409..

[7] A. Kipp, M. B. Wesenick, and F. Schiel, "Pronuncation modeling applied to automatic segmentation of spontaneous speech," *In EUROSPEECH-1997*, 1023-1026.

[8] K. Demuynck and T. Laureys, "A comparison of different approaches to automatic speech segmentation," *in Text, Speech and Dialogue, Springer Berlin Heidelberg,* 2002, pp. 277-284.

[9] Z.-H. Tan, and B. Lindberg, "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection," in *IEEE Journal of Selected Topics in Signal Processing*, 2010, vol. 4, no. 5, pp. 798 – 807, http://kom.aau.dk/~zt/online/rVAD/

[10] D. Vlaj, B. Kotnik, B. Horvat and Z. Kacic, "A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems," in *EURASIP Journal of Applied Signal Processing* 2005:4, 487-497.

[11] W. J. Hess, K. J. Kohler, and H. G. Tillmann, "The Phondat-verbmobil speech corpus," In *EUROSPEECH-1995*, 863-866.

[12] T. Bub, and J. Schwinn, "VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System", *Int. Conf. on Spoken Language Processing, 1996, Philadelphia, PA, USA*, vol. 4, pp. 2371-2374.

[13] Max Planck Institute for Psycholinguistics. (2001). WebCelex [database]. Retrieved from http://celex.mpi.nl/]

[14] I. Kraljevski, R. Kompe, R. Jäckel, G. Strecha, F. Kurnot, M. Rudolph, D. Hirschfeld, and R. Hoffmann., "Speech Quality Assessment in a Pronunciation Trainer for Speech Disorder Therapy", *In: Hoffmann, R. (ed.), Elektronische Sprachsignal-verarbeitung 2014: Tagungsband der 25. Konferenz Dresden, 26. - 28. März 2014, TUDpress, Dresden,* pp 153-160,.

[15] D. Huggins-Daines, et al, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. IEEE International Conference on. Vol. 1. IEEE*, 2006.

[16] Gstreamer [EB/OL], http://gstreamer.freedesktop.org/

[17] K. Sjölander, and J. Beskow, "Wavesurfer - an open source speech tool," *In ICSLP-2000,* vol.4, 464-467.

[18] X.L. Zhang, and J. Wu, "Deep belief networks based voice activity detection," in *Audio, Speech, and Language Processing, IEEE Transactions on* 21, no. 4 (2013): 697-710.