Voice Activity Detector (VAD) based on long-term phonetic features

Andrey Barabanov[1], Daniil Kocharov[2], Sergey Salishev[3], Pavel Skrelin[2], Mikhail Moiseev[4]
[1]Department of Cybernetics, Saint-Petersburg State University, Russia
[2]Department of Phonetics, Saint-Petersburg State University, Russia
[3]Department of Informatics, Saint-Petersburg State University, Russia
[4]Intel Labs, Intel Corporation, USA
https://doi.org/10.36505/ExLing-2016/07/0005/000264

Abstract
We propose a VAD using long-term phonetically motivated features with auditory masking, and pre-trained decision tree based classifier, which allows capturing syllable level structure of speech and discriminating it from common noise types. algorithm demonstrates on test dataset almost 100% acceptance of clear voice for English, Chinese, Russian, and Polish speech and 100% rejection of stationary noises independently of loudness with low computational cost.
Key words: Voice Activity Detector, classification, decision tree ensemble, auditory masking, phonetic features

Introduction
The problem of low complexity accurate VAD is important for many applications in Consumer Electronics, Wearables, Smart Home and other areas, where VAD serves as a low-power gatekeeper for a more complex and energy consuming Automatic Speech Recognition (ASR) system.

Our VAD approach is based on the detection of signal segments with formants in the spectrum. The method cuts off all voiceless consonants and the majority of voiced ones. This should be compensated by considering as a speech the sound signal that includes some unvoiced segments preceding and following vocalized sequence. The duration of such segments is language-dependent. On one hand, it should be long enough to contain consonant clusters. On the other hand, it should be shorter than inter-phrase pauses. Different languages have various consonant-to-vowel ratios and the maximum length of consonant clusters. Thus the length of consonant segment has to vary from language to language. The pause length is less language-dependent and more speaker-dependent. From this point of view the duration of consonant segment should be about 200 – 250 ms.

We propose to use long-term 200 ms speech statistics in combination with pre-trained complex non-linear classifier, which allows capturing syllable level structure of speech and distinguish it from common noises. Proposed algorithm substantially outperforms competitive solutions in various non-stationary noises and demonstrates on test dataset almost 100% acceptance of clear voice and 100% rejection of stationary noises at the cost of higher latency. The algorithm reuses short-term FFT analysis (STFFT) in ASR front-end; therefore, the complexity increase to MFCC ASR front-end is small.

VAD Algorithm Description

The algorithm consists of feature extraction, feature space dimensionality reduction and two-level classifier (phoneme and syllable levels). It uses Mel band spectral envelope and Mel band peak factor as features.

Spectral envelope is a standard ASR feature usually manifesting as MFCC or Linear Prediction Coefficients (LPC). According to the acoustic theory of speech production by (Fant, 1962), the harmonics of fundamental frequency contain most speech energy, which makes them robust to noise due to high SNR, it distinguishes speech from most types of noise. To improve noise robustness, tonal and temporal auditory masking are applied to spectral envelope (Fastl, Zwicker, 2006). Features are classified by a soft classifier using an ensemble of deep decision trees (Zhou, 2012). For classifier training we used database of continuous English speech TIMIT, noise databases Aurora 2, ETSI and SISEC10.

Comparison

For comparison, we used two state of the art VADs: Google WebRTC VAD and Nuance SREC VAD. For testing, we used sound files completely unused in training. We separately performed False Accept testing on noise database and False Reject testing on speech database with various SNRs. For false accept test, we used DEMAND database containing background noises for 18 environments (Table 1). We conclude that new VAD outperforms competitors. We tested false accept rate on 3 tracks of Rock, Pop, and Classic music genres not used in training. We conclude that new algorithm substantially outperforms competitors, still false accept on music is about 20%.

For false reject testing, we used speech database of 5 min recordings in four languages (English, Chinese, Russian, Polish – in accordance with their consonant coefficient), male and female speakers for each language with manual VAD markup. Noise was synthetically added to with various SNRs calculated as total speech to total noise power after high-

pass filter with 100 Hz cutoff. We conclude that new VAD is highly accurate and language and speaker insensitive for high SNR (up to 10 dB). We tested with various noises (Table 2).

Table 1. False accept rate comparison in % for different environmental noises and music.

Noise	SREC	WebRTC	Proposed
dkitchen	11.7	12.4	10.9
dliving	30.7	90.3	4.5
dwashing	20.9	84.5	5
nfield	0	74.1	0
npark	48.1	30	4.6
nriver	0.5	15.3	0
ohallway	23.4	15.4	2.7
omeeting	71.7	67.8	78.3
ooffice	28.6	20.8	0.3
pcafeter	77.4	80.9	38.3
presto	42.8	83.2	1.6
pstation	1	100	0
scafe	75.2	89.7	22.1
spsquare	31.4	73.3	11.8
straffic	10.6	82.9	0
tbus	67.1	77.2	30.4
tcar	1.2	95.5	0
tmetro	27.5	89.1	14.3
rock	91	97.6	11.1
pop	88.8	82.4	19.7
classic	91	90.7	18.1

New VAD algorithm is highly accurate in car noise with FAR about 1% at SNR 0 dB. For non-stationary noises, it demonstrates similar performance up to SNR 10 dB and degrades for lower SNR on babble noise. This correlates with subjective intelligibility of the speech.

Conclusion

The proposed algorithm substantially outperforms competitive solutions in various environments and demonstrates on test dataset almost 100% acceptance of clear voice and 100% rejection of stationary noises with 15% complexity increase compared to MFCC based ASR front-end. The algorithm has a latency of 200 ms, which is not acceptable for some scenarios such as VoIP. The algorithm in some cases falsely accepts some noises as voice: clatter of dishes; sound of flowing water; resonant

strokes; tonal beeps; babble noise; bird songs. The algorithm falsely rejects speech in the presence of high amplitude non-stationary noise especially babble noise.

Table 2. Proposed VAD false reject rate in % for different environmental noises and SNR.

language	gender	tcar			nriver			presto		
		20dB	10dB	0dB	20dB	10dB	0dB	20dB	10dB	0dB
English	f	0	0	0	0	1.6	0.7	0	25.6	54.7
	m	0.2	0.2	0.2	0.2	3.7	3.3	0.9	23.1	37.2
Chinese	f	0.1	0	0	0	1.4	5.1	0	27.6	40.8
	m	0	0	0	0	0.4	0.1	0	8.8	51.7
Russian	f	0.1	0.3	0	0.2	4	6.1	0.5	31.1	63.6
	m	0.2	0.2	0.3	0.2	3.1	3.1	0.1	20.4	65.8
Polish	f	0.5	0.7	0.5	0.6	5.1	5.8	1.6	41.1	53
	m	0.2	0.3	0.4	0.3	7.6	11.3	1.3	42.3	57.4
Mean		0.2	0.2	0.2	0.2	3.4	4.4	0.6	27.5	53

Table 3. False reject rate comparison in % for different environmental noises and SNR.

Noise	VAD	inf	20 dB	15 dB	10 dB	5 dB	0 dB
TCAR	SREC	0.9	1.3	1.6	1.9	2.2	2.5
	WebRTC	1.1	1.3	1.3	1.2	0.8	0.4
	Proposed	0.1	0.2	0.1	0.2	0.3	0.6
NRIVER	SREC	0.9	3.5	5.8	11.3	23.4	54.2
	WebRTC	1.1	3.1	4.7	7.1	12.0	22.8
	Proposed	0.1	0.2	0.8	3.4	9.8	27.5
PRESTO	SREC	0.9	2.0	2.6	4.3	9.2	21.4
	WebRTC	1.1	3.3	4.9	5.9	4.8	1.8
	Proposed	0.1	0.2	0.9	4.4	19.8	53.0

References

Fant, G. 1960. Acoustic theory of speech production: With calculations based on x-ray studies of Russian articulations.

Fastl, H., Zwicker, E. 2006. Psychoacoustics: facts and models, vol. 22. Springer Science & Business Media.

Zhou, Z.H. 2012. Ensemble methods: foundations and algorithms. CRC Press.