Journal of the Faculty of Engineering and Architecture of Gazi University 38:1 (2023) 497-508

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Investigation of factors affecting noise robustness in voice activity detectors

Selma Özaydın*
Department of Electronics & Communication Engineering, Engineering Faculty, Çankaya University, Etimesgut, 06790, Ankara, Türkiye

Highlights:

- Measurement of the factors affecting the noise-robustness of some voice activity detectors
- Comparative analysis of four different VAD detectors in adverse conditions
- Objective test results against the change in background SNR

Keywords:

- speech activity detection
- Speech analysis
- background noise
- Endpoint detection
- signal-to-noise ratio

Article Info:

Research Article
Received: 06.12.2020
Accepted: 12.02.2022

DOI:

10.17341/gazimmfd.836559

Correspondence:

Author: Selma Özaydın
e-mail: selmaozaydin@yahoo.com
phone: +90 312 233 1331

Graphical/Tabular Abstract

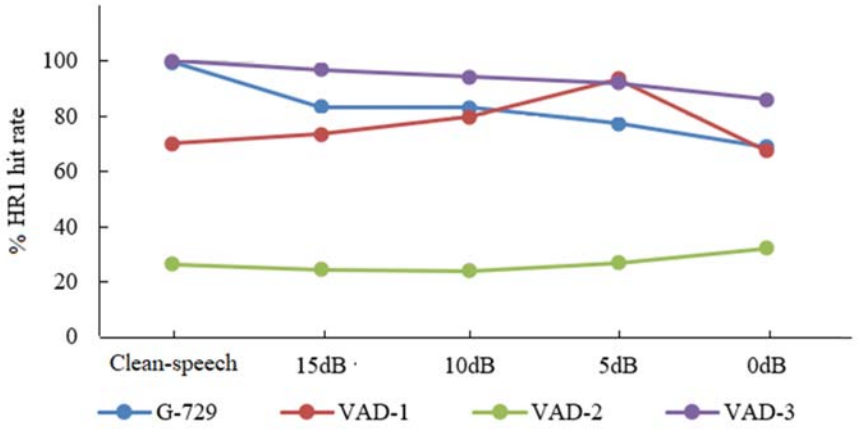


Figure A. HR1 hit rates

Purpose:
In the scope of this experimental study, some voice activity detectors (VADs) in the literature were tested to measure the changes in detection accuracy rates according to changing noise conditions. The main aim of this study was to investigate the factors affecting the robustness in varying acoustic noise conditions. In this context, the effect of situations in VAD methods such as whether the threshold value used in the decision phase is fixed or adaptive, the analysis window is short or long, the use of more than one feature vector together have been evaluated and analyzed comparatively.

Theory and Methods:
Four different VAD detectors in the literature were selected according to their feature vectors effecting the VAD decision algorithm such as short-term/log-term analysis or fixed/adaptive threshold usage. The efficiency of the detectors was tested on the NOIZEUS corpus to evaluate them under different acoustic conditions and to test them on a test data included in the literature. During the testing of the analyzed VADs, different types of input noise speech signals with environmental background noises between [15-0dB] such as restaurant, car, street, or station were tested. Speech hit rate (HR1) and silence hit rate (HR0) were measured.

Results:
The results were presented as figures of HR0 & HR1 rates against changing SNR conditions for each VAD detector. Besides, FEC, NDS, OVER, MSC accuracy rates were presented in a table to analyze each VAD performance in detail in adverse conditions. The results presented that the HR0/HR1 change rate of each detector were in different characteristics and there was no optimum detector result to represent a good performance both in HR0 & HR1 rates against changing acoustic SNR conditions. HR1 hit rates can be seen in Figure A.

Conclusion:
The main aim of this study was to investigate the factors affecting their robustness in varying acoustic noise conditions. From the test results it can be said that detection accuracy rates of VAD detectors according to changing noise conditions were affected from both short-term/long-term analysis and fixed/adaptive threshold usage. On the other hand, selected feature vectors seem to give chance to increase the performance of HR1 or HR0 rate. It seems difficult to design a detector with these feature vectors giving good performance both in HR0/HR1 rate.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:1 (2023) 497-508

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Konuşma aktivite detektörlerinde gürültülü dayanıklılığına etki eden faktörlerin incelenmesi

Selma Özaydın*
Çankaya Üniversitesi, Mühendislik Fakültesi, Elektronik & Haberleşme Mühendisliği Bölümü, Etimesgut, 06790, Ankara, Türkiye

Ö N E Ç I K A N L A R

- Bazı ses aktivitesi detektörlerindeki gürültü dayanıklılığını etkileyen faktörlerin ölçümü
- Dört farklı VAD detektörünün olumsuz koşullarda karşılaştırmalı analizi
- Arka plan SNR seviyesindeki değişikliğe karşı objektif test sonuçları

Makale Bilgileri
Araştırma Makalesi
Geliş: 05.12.2020
Kabul: 12.02.2022

DOI:
10.17341/gazimmfd.836559

Anahtar Kelimeler:
Konuşma aktivite algılama, konuşma analizi, arka plan gürültüsü; uç nokta tespiti, sinyal-gürültü oranı

ÖZ

Bu makalede, literatürdeki bazı konuşma aktivite detektörleri (VAD) farklı akustik gürültü koşullarında dayanıklılık performanslarına etki eden faktörler bakımından incelenmiş ve bu kapsamda değişen gürültü koşullarına göre doğru tespit oranlarındaki değişimleri test edilmiştir. Bu kapsamda, VAD metotlarındaki karar aşamasında kullanılan eşik değerin sabit veya uyarlamalı olması, analiz penceresinin kısa veya uzun olması, birden fazla özellik vektörünün birlikte kullanımı gibi durumların sonuç performansa etkisi değerlendirilmiş ve karşılaştırmalı olarak analiz edilmiştir. Bu makalede incelenen dört farklı VAD detektörünün üçü, karar sonucu üretirken kısa süreli analiz penceresi içerisindeki özellik vektörlerini kullanmakta iken, biri uzun vadeli spektral vektörlerin ölçüm sonucuna göre karar üretmektedir. Makale kapsamındaki VAD detektörleri, NOIZEUS gürültülü konuşma veri tabanı kullanılarak test edilmiştir. Bu sayede, analiz edilen VAD'ların performansı, literatürde halihazırda yer almış kapsamlı bir veri tabanı kullanılarak farklı akustik koşullar altında değerlendirilmiştir. Analiz edilen VAD'ların testi sırasında, restoran, araba, sokak veya istasyon gibi [15-0dB] arasında çevresel arka plan gürültülerine sahip farklı türde giriş gürültülü konuşma sinyalleri test edilmiştir. Testler objektif test ölçüm metotları kullanılarak yapılmış ve her bir VAD metodunun tespit doğruluk oranı ölçülmüştür. Sonuçlar, her bir yöntemin, olumsuz çevresel koşullarda farklı dayanıklılık performansı verdiğini göstermiştir.

Investigation of factors affecting noise robustness in voice activity detectors

H I G H L I G H T S

- Measurement of the factors affecting the noise-robustness of some voice activity detectors
- Comparative analysis of four different VAD detectors in adverse conditions
- Objective test results against the change in background SNR

Article Info
Research Article
Received: 06.12.2020
Accepted: 12.02.2022

DOI:
10.17341/gazimmfd.836559

Keywords:
Speech activity detection, speech analysis, background noise, endpoint detection, signal-to-noise ratio

ABSTRACT

In this manuscript, some voice activity detectors (VADs) in the literature were examined in terms of factors affecting their robustness under different acoustic noise conditions and in this context, the changes in detection accuracy rates according to changing noise conditions were tested. In this scope, the effect of situations such as whether the threshold value used in the decision phase in VAD methods is fixed or adaptive, the analysis window is short or long, the use of more than one feature vector together has been evaluated and analyzed comparatively. While three of the four different VAD detectors examined in this manuscript use feature vectors within the short-term analysis window while generating the decision result, one decides according to the measurement result of long-term spectral vectors. The VAD detectors in the article have been tested using the NOIZEUS noisy speech database. Thus, the performance of the analyzed VADs has been evaluated under different acoustic conditions using an extensive database that has already taken place in the literature. During the testing of the analyzed VADs, different input noise speech signals with environmental background noises between [15-0dB] such as restaurant, car, street, or station were tested. Tests were carried out using objective test measurement methods and the detection accuracy rate of each VAD method was measured. The results showed that each method gave different endurance performance in adverse environmental conditions.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *selmaozaydin@yahoo.com / Tel: +90 312 233 1331

1. Giriş (Introduction)

Sayısal konuşma işleme sistemleri, çeşitli türde akustik arka plan gürültüsü içeren konuşma sinyallerindeki, gerçekte konuşma olmayan bölümlerin konuşma olarak yanlış algılanmasını önlemek için genellikle bir VAD algoritmasına ihtiyaç duyarlar. Sayısal konuşma işleme uygulamalarında VAD algoritması kullanılarak konuşma aktivite bölgelerinin arka plan gürültüsünden ve konuşma olmayan sessiz bölgelerden ayrılması, analiz süresini en aza indirmekte ve hesaplama maliyetini azaltmaktadır. Ayrıca, bir VAD algoritması kullanılması durumunda, giriş sinyalindeki yalnızca konuşma olan bölümler işlenerek, konuşma işleme sisteminin daha doğru çalışmasına katkı sağlanır. Ayrıca, bir iletim hattından aktarılan konuşma işleme sisteminde, konuşma içermeyen sessiz bölgelerin VAD algoritması ile tespit edilip daha az bit sayısı kullanılarak tanımlanması sayesinde, sistemin toplam bit hızının azaltılması ve sistemin karşı tarafa iletilmesi için gerekli bant genişliğinden tasarruf edilmesi sağlanır [1]. Başarılı bir VAD algoritması, bir konuşma sinyalinin sınırlarını her türlü arka plan gürültüsünde sağlam bir şekilde etiketleyebilir. Bir VAD'ın doğruluğu, konuşma tabanlı bir uygulamanın genel performansını etkiler. Örneğin, bir konuşma tanıma sisteminin giriş sinyalinin başlangıç ve bitiş noktalarını yanlış algılaması, hatalı konuşma tanımaya yol açacaktır ve bu yüzden yüksek doğruluk oranına sahip bir VAD algoritması kullanılması önemlidir. Bir konuşma kodlama sisteminde VAD algoritması kullanılmasının amacı, sesli/sessiz bölgelerin tespit edilip sessiz bölgelerde daha az bit sayısı kullanılarak, bir ses sinyalini düşük bir veri hızıyla aktarabilmektir. VoIP içinde kullanılan bir VAD algoritması, konuşma olmayan bölgelerin iletiminden kaçınarak bant genişliği gereksinimini artırır. Tipik bir VAD algoritmasında sesli/sessiz veya sessizlik kararı, analiz çerçevesi içindeki sinyalden çıkarılan özellik vektörlerine göre bir karar modeline bakılarak verilir. Halihazırda olumsuz akustik koşullarda etkili bir şekilde çalışmak için tek bir özellik vektörü etkili olmadığından, literatürde önerilen birçok teknik birlikte kullanılarak karar modelleri oluşturulmaktadır [1, 2].

VAD algoritmaları, uç nokta (endpoint) algılama algoritmaları olarak da bilinir. Bir uç nokta algılama algoritmasının iki ana bölümü, ayırıcı akustik özelliklerin çıkarılması ve sonrasında bir dizi sınıflandırma kuralına dayalı (sesli/sessiz) kararıdır. Bir VAD algoritmasının en önemli sorunu, olumsuz akustik koşulların olduğu gürültülü ortamlarda konuşma içeren analiz çerçevelerinin yalnızca arka plan gürültüsü içeren analiz çerçevelerinden ayrılmasıdır. Konuşma işlemede VAD algoritmaları için, zaman düzleminde veya frekans düzleminde işletilebilen ve etkinlikleri doğruluk yüzdeleri ile hesaplama maliyetlerine göre değerlendirilebilen pek çok farklı yaklaşım önerilmiştir. Literatürde sıfır geçiş oranı (ZCR) ve/veya sinyal enerjisi [3-6], perde frekansı [7-10], otokorelasyon [11-13], spektrum [14-16], kepstrum [17-20], dalgacıklar [21, 22], ana bileşen analizi [23] veya istatistiksel model tabanlı algoritmalar [24-27] gibi özellik vektörlerini kullanan birçok algoritma türü vardır. Bu yöntemlerin çoğu gürültülü konuşma sinyalindeki SNR (sinyal-gürültü oranı) değişimlerine duyarlıdır ve gürültünün belirli zaman aralıklarında sabit olduğunu varsayarlar. Son zamanlarda, denetimli (supervised) modeller kullanıldığındaki büyük miktarlarda etiketlenmiş eğitim verisi ihtiyacının üstesinden gelmek için denetimsiz (unsupervised) modeller önerilmiştir [28-31]. Telefon ve multimedya iletişimindeki VAD uygulaması için, ITU-T VAD standardı G.729B VAD geliştirilmiştir [3]. Mobil iletişim sistemleri için diğer uygulama, ETSI tarafından standart hale getirilen, iki ayrı versiyonu bulunan değişken bir hızlı AMR (Adaptive Multi Rate) kodlayıcıdır. Her iki sürüm (versiyon) için de giriş sinyali alt frekans bantlarına bölünmekte, alt bantlarda yapılan analiz sonuçları ile seçilen diğer özellik parametrelerinin ortak değerlendirilmesi sonucu VAD kararı üretilmektedir [32-34].

Zaman düzlemindeki VAD algoritmaları, frekans düzlemindeki VAD algoritmaları ile kıyaslandığında hesaplama karmaşaları daha azdır ve böylece zaman gecikmeleri minimum seviyededir. Sinyal enerjisi hesabına dayalı zaman düzlemi VAD algoritmaları, konuşma olan bölgelerin enerjisinin arka plan gürültü içeren bölgelerden nispeten daha yüksek olduğunu varsayarak, her analiz penceresindeki enerjiyi tahmin eder. Algoritmanın başında sabit bir eşik değer için, bir enerji eşik değeri tanımlanır ve konuşmanın başlangıç ve bitiş noktalarının tanımında kullanılır. Bu nedenle, bir analiz penceresindeki konuşmanın enerji genliği, analiz penceresindeki sinyali konuşma etkin veya etkin değil (sessiz) olarak sınıflandırmak için önemli bir parametredir. Enerji tabanlı VAD algoritmaları düşük hesaplama karmaşası sebebiyle basitlikten yararlanırken, diğer yandan da arka plan gürültüsüne çok duyarlıdırlar. Bu nedenle, konuşmanın başlayıp bittiği uç noktaları doğru tanımlamak için gürültüye dayanıklı algoritmalar gereklidir [35, 36].

Bir konuşma sinyalinin VAD bölgelerini değerlendirmek için kısa süreli enerji metodu kullanıldığında, özellikle arka plan gürültülerin varlığında, kısa süreli analiz penceresi içindeki sinyalin enerji değerini sabit bir eşik değerle kıyaslayıp karar üreten zaman düzlemindeki enerji tabanlı VAD metotları yetersiz kalite sağlar [3, 37]. Ayrıca, insan işitsel sisteminin, yüksek genlikli seslerin düşük genlikli seslerle aynı çözünürlüğü gerektirmediği logaritmik bir sürece sahip olduğuna inanılmaktadır. Bu nedenle, bir konuşmada doğru bilgileri elde etmek için konuşma olmayan bölgelerin doğru tanımı çok önemlidir. Sonuç olarak, bazı VAD algılama algoritmaları için, her bir ifadedeki başlangıç/bitiş noktalarını tam olarak saptamak için bir ileri/geri arama algoritması uygulamak gerekebilir [1, 2]. İleri/geri arama algoritmaları ise, zaman gecikmesine ihtiyaç duyduklarından konuşmanın gerçek zamanlı işlenmesi için uygun değildir. Tek tip enerji hesaplaması yerine, sinyalin genliğine göre ayarlanmış tek tip olmayan bir hesaplama, küçük genlikli sinyallerinin tespitini geliştirebilir. Sinyal sıkıştırma parametresi olarak kullanılan bir μ değerine ayarlanmış logaritmik bir ölçeğin uygulanması, seçilen μ değerine bağlı olarak tepe genliklerini bastırırken düşük genlik sinyallerini artırmaktadır [38].

Bir VAD algoritması için performans parametreleri; basitlik, arka plan gürültüsüne karşı sağlamlık ve gürültülü ortamlarda bile sözcük sınırlarının tam olarak tespiti olabilir. Bu nedenle, bir VAD algoritması gecikme, hassasiyet ve doğruluk açısından değerlendirilmektedir. Bu gereksinimleri karşılamak için, bu makalede incelenen VAD algoritmaları, konuşmada ifade içeren bölümleri bulmak için genellikle zaman düzleminde çalışmakta ve enerji bazlı ölçüm yöntemleri kullanılmaktadır. İncelenen VAD'ların kalitesi, tespit oranı, arka plan gürültüsüne karşı sağlamlığı ve düşük hesaplama karmaşıklığı değerlendirilerek ölçülmüştür.

Bu makale aşağıdaki gibi düzenlenmiştir. İkinci bölümde önceki çalışmaların bir incelemesi ve bir VAD algoritmasının genel bir tanımı sunulmaktadır. Üçüncü bölümde, incelenen VAD algoritmaları tanıtılmakta ve uygulanan test metotları hakkında teorik bir bilgi vermektedir. Dördüncü bölümde, seçilen VAD yöntemlerini değerlendirmek için yapılan testlerin sonuçları sunulmaktadır. Son bölüm, test sonuçları kapsamındaki yöntemleri değerlendirerek makaleyi sonlandırmaktadır.

2. Literatür Taraması: İlgili Çalışmalar ve Motivasyon (A Literature Review: Related Works and Motivation)

Literatürde VAD detektörü ile ilgili son yıllardaki çalışmalar incelendiğinde, önemli bir kısmının spektrum analizine dayalı veya derin öğrenme tabanlı detektörler üzerinde yoğunlaştığı görülmekte ve performans bakımından incelendiklerinde sonuçlarının değişik

akustik gürültülü şartlarında oldukça başarılı olduğu görülmektedir. Diğer yandan spektrum analizine dayalı ve derin öğrenme yaklaşımlı metotların, özellikle zaman düzlemindeki özellik vektörlerini kullanan metotlarla kıyaslandığında, hesaplama karmaşalarının daha yüksek olduğu ve karar süreçlerinin göreceli yavaşlığı dikkate alınması gereken bir husustur.

Bu bölümde, son yıllarda önerilen bazı VAD detektörleri, kullandıkları metotlar ve performans test sonuçları bakımından incelenmiştir. Bu kapsamda, algoritmalarında kullandıkları akustik öznitelik vektörleri, eğitim ve test aşamasında seçtikleri veri tabanları, performans ölçüm metotları ve sınıflayıcı türleri incelenmiştir. Kimi çalışmalarda tek bir öznitelik vektörü kullanılırken, kimi çalışmalarda VAD karar aşamasında birden fazla öznitelik vektörü birbirini tamamlayıcı nitelikte kullanılmıştır. Böylece bir yöntemde oluşabilecek tespit hatasının, başka bir yöntemle denetimi sonucunda performansta iyileştirme başarılmaktadır. Diğer yandan, kullanılan öznitelik vektörlerinde, seçilen veri tabanlarında ve performans ölçüm metotlarındaki çeşitlilikler sebebiyle, performans verileri bakımından birebir kıyaslama yapılamamıştır. Bu bakımdan, VAD detektörlerinin gerçek performans kıyaslamasının sadece burada sunulan performans verilerine bakılarak değil, ilgili makale ele alınıp yöntemlerinin detayları incelenerek ve çeşitli testler altında ortaya koydukları performans değerlerine bakılarak ortaya konabileceği değerlendirilmektedir.

Yoo vd. [35], sesin formant frekanslarının (sesin spektrumundaki spektral tepeler) gürültüye karşı dayanıklılık özelliklerini kullanarak, çeşitli gürültü türlerine karşı dayanıklı olduğunu belirttikleri ve düşük SNR koşulları altında çalışabilecek bir yöntem önermektedir. Konuşma sinyalinde, enerji belirli spektral bantta yüksek oranda yoğunlaştığında bir spektral tepe (formant) oluşmaktadır. Spektral tepelerin önemi, gürültü kaynaklı ciddi bozulmalardan sonra bile hayatta kalması muhtemel olmasıdır. Ancak halihazırda, yüksek gürültülü bir konuşma sinyalinde, gürültü sebebiyle oluşan gürültü kaynaklı spektral tepelerden sinyalin kendi spektral tepelerini doğru bir şekilde çıkarmak oldukça zor bir problemdir. Bu nedenle, makalede, gürültülü koşullarda spektral zirveleri tespit edebilmek için, spektral tepe tabanlı yeni bir VAD algoritması önerilmektedir. Önerilen metot şu varsayıma dayanmaktadır; Spektral tepe noktalarının, belirli frekanslar içinde çevreleyen frekanslardan çok daha yüksek enerjilere sahip olduğu ve belirli bir spektral tepe noktasının varlığının, çevreleyen bölgelerdekine göre tepe bölgesindeki enerjinin göreceli farkını ölçülerek belirlenebileceği önerilmektedir. Bu nedenle, komşusundan önemli ölçüde daha yüksek bir enerjiye sahip bir spektral tepe mevcutsa, ilgili olmayan spektral zirvelerin etkisini azaltılmaktadır. Bu amaçla, İngilizcedeki sesli sesleri temsilen 8 adet spektral tepe şablonu kullanılmaktadır. Önerilen metot, TIMIT veri tabanı kullanılarak ve 8 farklı gürültü türü ile çeşitli SNR koşulları (0-30dB arası) altında performans bakımından değerlendirilmiştir. Deneylerde örneğin 10dB ortam gürültü sinyali koşullarında G729.B VAD detektörü ile kıyaslandığında ortalama %50 civarı, 0dB ortam gürültülü giriş sinyali için ise %52 civarında tespit doğruluğunun arttığı görülmüştür.

Son yıllarda önerilen VAD detektör yaklaşımlarından öne çıkan diğer bir metot Muralishankar vd. tarafından önerilmiştir [39]. Çalışmada, VAD tespitindeki iyileştirme amacıyla Mel frekansı kepstral katsayılarındaki (MFCC) geleneksel üçgen filtre yaklaşımına [40] alternatif bir yöntem önerilmektedir. Bu kapsamda, her bir filtre cevabının bitişik komşusuyla üst üste binmesini sınırlayan yeni bir değiştirilmiş Mel-kesikli kosinüs dönüşümü (MMD) filtre bankası yapısı önerilmiştir. MFCC katsayılarının çıkarılmasında halihazırda kullanılan üçgen filtrelerin aksine, önerilen MMD filtre yapısının daha yumuşak bir tepkiye sahip olduğu ve tek bir işlemde ayrı kosinüs dönüşümü ve Mel ölçekli filtreleme sunabildiği belirtilmektedir. Normalde VAD tespiti için tek özellik vektörü olarak MFCC seçildiğinde performans artışında önemli kazanım sağlanamadığı halde, önerilen çalışmada ses sinyalinin uzun vadeli diferansiyel entropisinin izlediği MMD filtre bankaları kullanarak, VAD tespit doğruluğunda önemli iyileşme elde edildiği belirtilmektedir. Testlerde Switchboard veri tabanı kullanılmış, NOISEX-92 gürültü verisi eklenerek gürültülü ortam testleri gerçekleştirilmiştir. MMD tabanlı VAD detektörünün performans testlerinde G729.B ile kıyaslamada, örneğin 0dB ortam gürültüsü şartlarında HR1 tespit oranı bakımından %15 civarı, HR0 tespit oranı bakımından ise performansta %40 civarında artış elde edilmiştir. Ayrıca, söz konusu çalışmada test edilen G.729B sonuçları ile bu makalede Tablo 1'deki G.729B sonuçları 0dB gürültü koşullarında kıyaslandığında, HR1 ve HR0 test sonuçlarının birbirine yakın olduğu görülmüştür.

Derin sinir ağları tabanlı olarak son yıllarda önerilen bazı yaklaşımlar ise şu şekildedir; Sehgal ve Kehtarnavaz çalışmalarında [41], evrişimli sinir ağına dayalı gerçek zamanlı ses etkinliği algılaması yapan bir akıllı telefon uygulamasını sunmaktadır. Evrişimsel sinir ağları VAD detektörü olarak kullanıldığında, karar süresindeki yavaşlık gerçek zaman uygulamalarını güçleştirmektedir. Makalede de bahsedildiği üzere, derin öğrenme yaklaşımları, VAD tespitini daha etkili bir şekilde gerçekleştirilebildikleri halde, bu tür yaklaşımlar çok uzun karar çıkarım sürelerine sahiptir ve bu da gerçek zamanlı konuşma işleme kullanımlarına engel teşkil etmektedir. Bunun temel nedeni, sinir ağı mimarilerinin oldukça büyük ve derin olarak tanımlanmasıdır. Önerilen çalışmada ise, akıllı telefon uygulamalarında gerçek zamanlı çalışabilecek ve gürültüye dayanıklı bir algoritma önerilmektedir. Yöntem olarak, gerçek zamanlı bir VAD tespiti için pratik bir CNN mimarisinin geliştirildiği belirtilmektedir. Bu amaçla kullanılan teknikler kısaca şöyledir: Kısa süreli pencereler içine alınan konuşma sinyali, öncelikle logaritmik Mel-ölçekli enerji spektrum görüntüleri olarak temsil edilmektedir. Akabinde CNN'lere giriş olarak uygulanmaktadır. Bu şekilde, kısa süreli Fourier Dönüşümüne (STFT) göre daha az katsayıya sahip olduğu ve daha kısa bir çıkarım süresine, sonuçta daha küçük CNN mimarisine yol açtığı belirtilmektedir. 15 farklı arka plan gürültüsü içeren DCASE2017 veri tabanı kullanılarak gerçekleştirilen deneysel çalışmada ölçüm kriteri, HR1 ve HR0 tespit oranları bakımından olmuştur. Örneğin, 0dB SNR seviyesinde ofis gürültülü giriş sinyali için CNN detektör'ün HR1 tespit performans değerinde, G729.B standart VAD detektörüne göre %15 civarı artış görülmektedir.

Sesli/sessiz karar sınıflandırmasını belirlemek için Ivry vd. [42], VAD tespitinde derin öğrenme tabanlı bir yaklaşım sunmuşlardır. Söz konusu VAD detektörü, difüzyon haritaları yöntemi uygulanarak oluşturulan, spektral özellikleri zamansal bilgilerle düşük boyutlu temsilleriyle eşleştiren bir kodlayıcı ve kod çözücü içermektedir. Önerilen metotta, difüzyon haritaları yöntemi kodlayıcı-kod çözücü tabanlı iki bağımsız eğitilmiş derin sinir ağı ile entegre edilerek, konuşma olan ve konuşma dışı pencereler ayrı ayrı temsil edilmiştir. Yöntemde, sabit olmayan sesler yüksek değişkenlik gösteren doğaları nedeniyle hatalı VAD tespitinin sebebi olarak gösterilmiştir. VAD detektörünü eğitmek için, bir görüntü-ses birlikte toplanmış veri tabanındaki ses kayıtları kullanılmıştır. Test sonuçları, konuşma olan ve olmayan bölgelerin toplam tespit doğruluğu cinsinden ölçülmüş ve örneğin 10dB ortam gürültüsü içeren giriş sesleri için %99 civarı başarı elde edilmiştir. Önerilen modelin gerçek zamanlı çalışabildiği ifade edilmektedir.

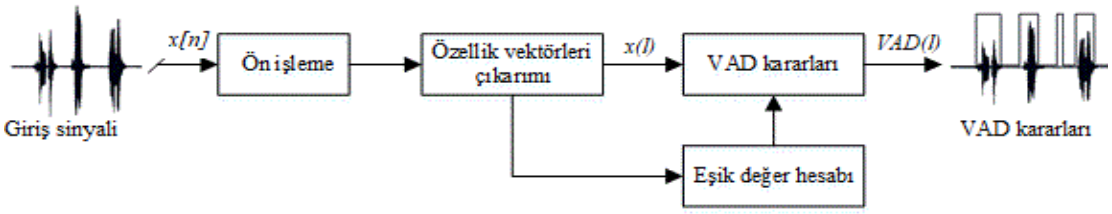
Akustik gürültülü ortamlara dayanıklı bir VAD detektörü için, Ali ve Talha [43], çalışmalarında referans verdikleri Katz metodunu kullanarak hesapladıkları uzun süreli pencere aralıklarında çıkarılan özellik vektörlerini kullanmışlardır. Önerilen yöntem, Arapça King Saud Üniversitesi (KSU) veri tabanı ve Texas Instruments Massachusetts Teknoloji Enstitüsü (TIMIT) veri tabanından seçilerek

oluşturulan İngilizce veri tabanı olarak iki farklı dildeki veri tabanı ile test edilmiştir. Yöntemin değerlendirilmesi, konuşma içeren ve konuşma içermeyen bölgelerin doğru veya yanlış tespit oranları cinsinden oluşturulan ve makalede önerilen bir doğruluk yüzdesi formülü cinsinden ölçülmüştür. Önerilen ölçüm metodu ile, TIMIT veri tabanı için örneğin 5dB gürültü seviyesinde yaklaşık %85, KSU veri tabanı için ise yaklaşık %88 doğruluk oranında tespit başarıldığı belirtilmektedir.

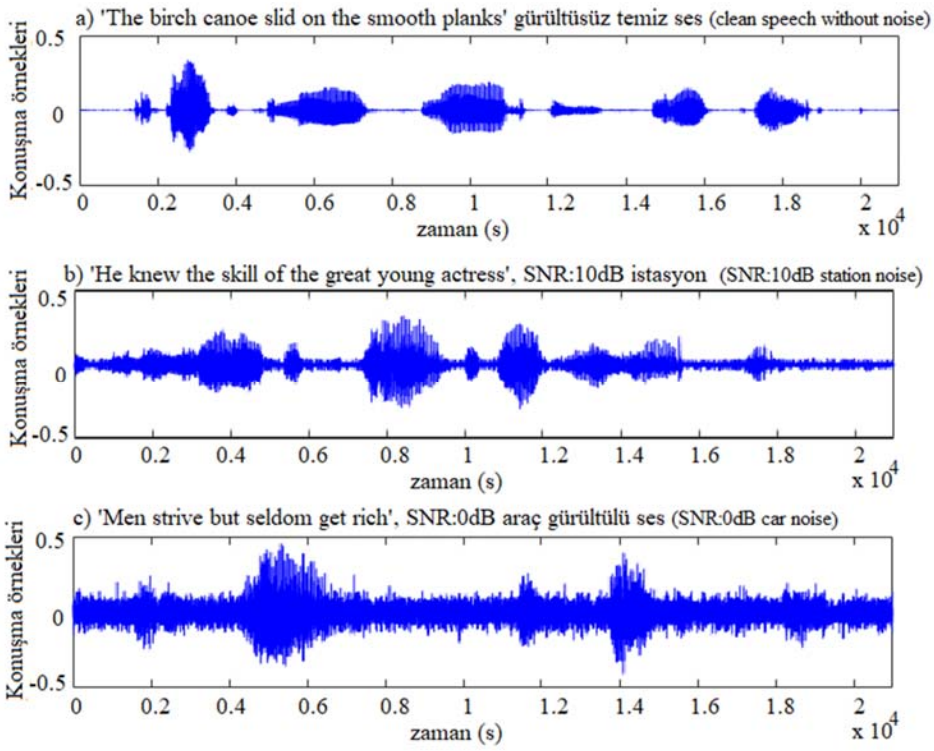
Lee vd. tarafından yapılan araştırmada ise [44], gürültüye dayanıklı bir VAD sistemi yaklaşımı önerilmektedir. Önerilen yöntem, spektral ve zamansal davranışı kullanarak derin öğrenmeye dayalı bir mekanizma uygulamaktadır. Önerilen yöntem, bilinen veya bilinmeyen gürültü eklenmiş deneylerde ve gerçek zamanlı gürültülü verilerle, diğer birkaç derin öğrenme temelli yöntemle karşılaştırılmıştır. Sonuçlar, önerilen yöntemin, dikkate alınan tüm senaryolarda diğer yöntemlerden daha iyi performans gösterdiğini, bilinmeyen veya beklenmedik gürültü ortamlarında da iyi bir davranış sergilediğini göstermektedir. Eğitim aşamasında TIMIT veri tabanı kullanılmış, değişik seviyelerde 8 farklı SNR gürültüsü için NOISEX-92 gürültü verisi eklenmiştir. Testlerde ise LibriSpeech veri tabanı kullanılmış ve 4 farklı test senaryosu gerçekleştirilmiştir. Metot, derin öğrenme tabanlı diğer metotlarla karşılaştırılmıştır. Performans testlerinde, örneğin 0dB ortam gürültüsünde %92 civarı başarı elde edilmiştir.

3. Değerlendirilen Vad Metotları ve Uygulanan Test Metotları
(Evaluated Vad Methods and Applied Test Methods)

Bir VAD algoritmasının temel prensibi Şekil 1'de görülmektedir. Bu makalede VAD detektörlerinin değerlendirilmesinde kullanılan NOISEUS veri-kümesindeki temiz ve akustik gürültülü giriş konuşma sinyallerinden bazı örnekler ise Şekil 2'de görülmektedir. Buradaki temel prensip, giriş sinyali üzerinde bir ön işleme sürecinin ardından, her bir analiz penceresinde VAD kararı için kullanılacak özellik vektörlerinin çıkarılması, akabinde belirlenen eşik değere göre konuşma olan bölümlerin konuşma içermeyen sessiz (veya gürültü içeren) bölgelerden ayrılmasıdır. VAD algoritmaları, ön işleme sırasında analiz pencerelerine ayrılan konuşma sinyali parçaları üzerinde işlemlerini gerçekleştirerek, her bir analiz penceresi için karar aşamasında, kıyaslanan eşik değer üzerinde bir değer ortaya çıkarsa 'sesli' (VAD=1), diğer durumda 'sessiz' (VAD=0) şeklinde iki durumlu bir sonuç üretirler. Genel bir VAD detektörü akış diyagramı Şekil 3'te görülmektedir. Konuşma olmayan bölümler,

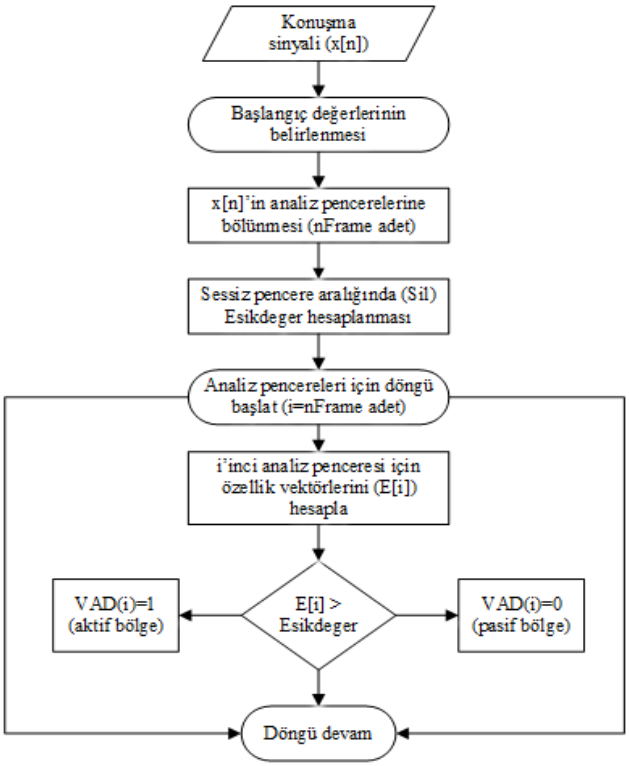


Şekil 1. Genel bir VAD algoritmasının blok şeması (Block diagram of a general VAD algorithm)



Şekil 2. NOISEUS veri tabanındaki konuşma sinyalindeki bazı cümle örnekleri
(Some examples of sentences in the speech signal in the NOISEUS database)

çeşitli VAD algoritmalarında gürültü olarak ta adlandırılmaktadır. Analiz pencere uzunluğu her algoritmada farklılık göstermektedir ve 5-40ms aralığında değişmektedir. VAD algoritmalarının doğruluğu ve güvenilirliği, uygulanan metotlar yanında seçilen eşik değere de bağlıdır. Bazı VAD uygulamalarında eşik değer sabit iken; diğer bazı VAD algoritmaları taban gürültüsüne göre eşik değeri güncellemektedir.



Şekil 3. Genel bir VAD metodu akış diyagramı
(Flowchart of a general VAD method)

Enerji tabanlı VAD algoritmalarında eşik değer hesabı için genel yaklaşım şu şekildedir; Eş. 1'deki gibi, başlangıçta belli bir sürede (v adet analiz penceresi) sinyal içerisinde konuşma olmadığı varsayılarak, seçilen analiz pencereleri içindeki sinyallerin ortalama enerjisi (E_r) hesaplanmaktadır. Daha sonra, herhangi bir analiz penceresindeki sinyal enerjisi (E_j), E_r ile kıyaslanmakta ve Eş. 2'deki senaryoya göre karar verilmektedir (burada *k* çarpanı, güvenilir bir eşik değer belirleyebilmek için kullanılan sabit bir değerdir). Bunun yanında, konuşma olmayan pencerelerdeki arka plan gürültüsüne göre eşik değerini sürekli olarak uyarlayan algoritmalarda bulunmaktadır.

$$E_r = \frac{1}{v} \cdot \sum_{m=0}^{v} E_m \tag{1}$$

$$VADkarar\imath: \begin{cases} 1, & eğer\ (E_j > k.E_r) \\ 0, & diğer\ durumlar\ için \end{cases} \quad (k > 1) \tag{2}$$

Genel olarak, j'inci analiz penceresi içerisindeki N-örnek sayısına sahip bir ses sinyalinin *i*'inci örneği x(i) ise, analiz penceresi f_i, Eş. 3'te olduğu gibi ile temsil edilebilir.

$$f_i = \{x(i)\}_{i=(j-1).N+1}^{j.N} \tag{3}$$

Bu makalede ele alınan VAD algoritmalarının özellikleri aşağıdaki alt başlıklarda özetlenmiştir. Bu kapsamda, kıyaslama amaçlı değerlendirilen VAD2 ve G.729B VAD algoritmaları 10ms analiz pencerelerinde VAD kararları üretirken, VAD1 algoritması 50ms analiz pencerelerinde VAD kararları üretmektedir. 8kHz'de örneklenmiş test veri-tabanı için, 10ms analiz penceresi, 80 örnek sayısına denk gelmektedir. Ayrıca VAD1 sabit bir eşik değer kullanmakta iken, G.729B, VAD2 ve VAD3 algoritmalarında kullanılan eşik değer değişen akustik koşullara göre uyarlanmaktadır. VAD3 detektörü, diğerlerinden farklı olarak uzun analiz penceresi aralığında çıkarılan özellik vektörlerini kullanarak karar üretmektedir.

3.1. VAD1 Metodu (VAD1 Method)

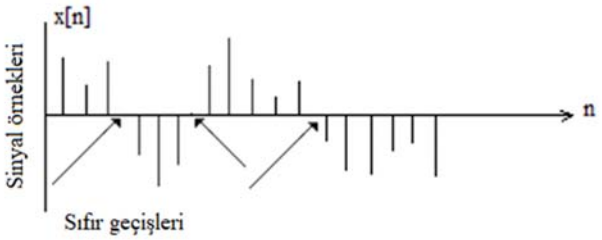
Bachu vd. tarafından yapılan çalışmada [37] enerji ve ZCR kullanılarak VAD kararı veren algoritma üzerinde durulmuştur. Öncelikle giriş konuşma sinyalleri üst üste binmeyen bölümlere bölünerek, her bölümün enerji ve ZCR değerleri hesaplanmış, konuşmanın en başında sessiz olarak kabul edilen bir bölgeden enerji ve ZCR için eşik değerler belirlenmiş, VAD kararının verilmekte zorlandığı bölgelerde karar iyileştirmesi için konuşma bölümleri tekrar iki bölüme ayrılarak bu bölgelerde analiz tekrarlanmıştır. Önerilen metot gürültüsüz giriş sinyalleri için denenmiş, gürültülü sinyallere karşı etkinliği ileride gerçekleştirilecek bir araştırma olarak belirtilmiştir. VAD1 metodu [37], sesli/sessiz kararı için ZCR ve enerji hesabını birlikte değerlendirerek analiz yapmaktadır. Zaman düzleminde analiz penceresi içindeki sinyal örneklerinin (x(n)) enerji hesabı, Eş. 4'e göre ve ZCR hesabı Eş. 5'e göre yapılmaktadır. Şekil 4'te görüldüğü üzere, herhangi bir kesikli zaman x[n] sinyalinin ZCR değeri, x[n] sinyal örneklerinin yatay eksene göre işaret değiştirme miktarına göre belirlenmektedir. Burada E_n, hamming penceresi (w(n)) kullanılarak pencerelenmiş *n*'inci analiz penceresinin ortalama enerjisini göstermektedir.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m).w(n-m)]^2 \tag{4}$$

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|.w(n-m) \tag{5}$$

burada $sgn[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ 0, x(n) < 0 \end{cases}$

ve $w(n) = \begin{cases} 0{,}54 - 0{,}46.Cos(\frac{2\pi n}{N-1}), & 0 \leq n \leq N-1 \\ 0, diğer\ durumlar \end{cases}$



Şekil 4. Sıfır geçiş oranı (ZCR) genel tanımı
(Zero crossing rate (ZCR) general description)

VAD1 algoritmasında öncelikle, giriş konuşma sinyalinin ilk 10 analiz penceresinde konuşma olmadığı varsayılarak, arka plandaki gürültü sinyali enerjisinin ortalama değeri ölçülmekte (E_r, Eş. 1) ve Eş. 2'deki gibi sabit bir *k* değeri seçilerek E_{min} belirlenmektedir. Akabinde, başlangıçta belirlenen bir λ sabit değeri kullanılarak enerji eşiği sabit değeri (ITL) Eş. 8'e göre hesaplanmaktadır. ITL hesabında ihtiyaç duyulan maksimum ve minimum enerji değerleri için, tüm giriş sinyalinin analiz pencerelerindeki enerji değerleri hesaplanmaktadır. Bu şekildeki ITL hesaplama metodu ise, VAD1 metodunun gerçek zamanlı işletilmesine imkan vermemektedir. ITL değeri belirlendikten sonra, *j*'inci analiz penceresindeki N örnek sayısına sahip giriş sinyali x(n) için analiz penceresi toplam enerjisi (E_n), Eş. 4'teki genlik kare enerji metodu ile hesaplanmaktadır. ITL

değerinin üstünde kalan enerji değerleri (E_n) ve Eş. 5'e göre hesaplanan ZCR değerine (Z_n) bakılarak, konuşma sinyalinin var olduğu bölgeler belirlenmektedir. Belirlenen limit değerler için karar güçlüğü çekilen durumlarda, analiz penceresi ikiye bölünerek, alt analiz pencerelerinde tekrarlanan algoritma ile, karar süreci iyileştirilmeye çalışılmaktadır.

3.2. VAD2 Metodu (VAD2 Method)

VAD2 metodu [48], arka plan gürültüsüne göre uyarlamalı bir eşik değer belirlenmesine dayanmaktadır. Eşik değer (E_{deger}) Eş. 6'ya göre belirlenmektedir.

$$E_{deger} = (1-\lambda).E_{max} + \lambda.E_{min} \quad (6)$$

Burada E_{deger} uyarlanan eşik değer, E_{max} ve E_{min}, önceki analiz pencerelerinde hesaplanarak sürekli güncellenen maksimum ve minimum enerji değerleridir. λ değeri, optimum bir performans için eşik değeri interpolasyon işlemine tabii tutabilmek için kullanılan katsayıdır ve Eş. 7'ye göre tanımlanmıştır. Yine VAD1 metodundan farklı olarak analiz penceresi enerji hesabında Eş. 8'deki RMSE (ortalama genlik karelerin kökü hatası - root mean square error) enerji formülü kullanılmaktadır. Taban gürültüsüne göre sürekli güncellenen eşik değer hesabı, gürültülü koşullardaki VAD kararı sonucunu etkilemektedir.

$$\lambda = \frac{E_{max}-E_{min}}{E_{max}} \quad (7)$$

$$E_j = \left[\frac{1}{N}.\sum_{i=(j-1).N+1}^{j.N} x^2(i)\right]^{\frac{1}{2}} \quad (8)$$

3.3. G.729B VAD Metodu (G.729B Method)

G.729B VAD [32, 45, 46], ITU-T tarafından, sabit telefon ve çoklu medya iletişimleri için standart olarak kabul edilmiş bir VAD kodlayıcıdır ve analiz penceresi 10 ms olarak belirlenmiştir. 8000Hz'de örneklenmiş bir ses sinyali için bu 80 örnek sayısına tekabül etmektedir. G.729B VAD algoritmasında özellik vektörleri olarak, 0-1kHz bant aralığındaki diferansiyel güç hesabı, tüm bant diferansiyel güç hesabı, çizgi spektrum katsayıları (LSF) ve sıfır geçiş oranı (ZCR) gibi dört ana parametreye bakılarak VAD kararı verilmektedir.
G.729B kodlayıcı, büyük veri tabanları kullanılarak farklı akustik gürültülü koşullarda test edilmiş ve değişik SNR seviyelerinde VAD performans ölçümleri yapılmıştır. Testler sonucunda, 15dB'ye kadar olan gürültülü koşullarda başarılı sonuçlar elde edilmiştir [32, 45, 46]. Ancak, G.729B detektörünün yüksek gürültülü sinyaller için performansı düşük olarak değerlendirilmektedir.

3.4. VAD3 Metodu (VAD3 Method)

VAD3 detektörü, gürültülü ortamlarda konuşma algılama sağlamlığını ve konuşma tanıma sistemlerinin performansını iyileştirmek için önerilmiştir ve algoritma, konuşma ve gürültü arasındaki uzun vadeli spektral sapmayı (LTSD) ölçmektedir. Uzun vadeli spektral zarf ve ortalama gürültü spektrumu karşılaştırılarak, konuşma/konuşma dışı karar kuralını formüle etmektedir. Böylece karar algoritmasının iyileştirildiği belirtilmektedir. Önerilen metot, VAD2'ye benzer şekilde gürültü enerjisine uyarlamalı bir eşik değer kullanmaktadır. Karar eşiği, ölçülen gürültü enerjisine uyarlanmaktadır. Yapılan testlerde, 6 adet analiz penceresi üzerinden LTSD analizinin konuşmayı ve gürültüyü sınıflandırmada optimum sonucu verdiği gözlemlenmiştir. Önerilen algoritma, konuşma/sessizlik ayrımı açısından, sahada sık kullanılan bazı VAD'larla karşılaştırılmış ve deneysel sonuçlar, referans olarak kullanılan G.729B ve uyarlanabilir çoklu oran (AMR) gibi standart VAD'lara ve dağıtılmış konuşma tanıma için gelişmiş ön uç VAD'larına göre bir avantaj gösterdiğini ortaya koymuştur [2].

3.5. Uygulanan Nesnel Test Metotları (Applied Test Methods)

VAD metotlarının performans ölçümlerinin yapılmasında literatürde yaygın olarak, aşağıda tarif edilen 4 farklı nesnel ölçüm parametresi kullanılmaktadır ve her bir parametre aşağıda açıklanan farklı hataları karakterize etmektedir. Kıyaslama için kullanılan temiz konuşma kayıtları için VAD kararları ise genellikle el ile işaretlenerek sesli/sessiz (veya arka plan gürültüsü olan sinyallerde gürültü) bölgeler belirlenmektedir [47]

3.5.1. Ön uç kırpma (FEC-Front end clipping)

FEC, sessiz bölge bitip yeni bir ifade başlarken, ifadenin başlangıcını yanlışlıkla hala 'sessiz' olarak sınıflandıran hataları, toplam sesli örnek sayısı cinsinden, yüzdelik oranda ölçer (Eş.9). Eş. 9'da, N_F, orijinal konuşma sessiz'den sesli'ye geçerken, test edilen VAD metodu tarafından 'sessiz' olarak tanımlanan giriş sinyal örneklerinin sayısı, N_{speech}, toplam sesli örneklerin sayısıdır.

$$FEC = \frac{N_F}{N_{speech}} x100 \quad (9)$$

3.5.2. Taşma hatası (OVER-Over error)

OVER, bir ifade bitip sessiz bölge başlamışken, ilgili VAD metodu tarafından yanlışlıkla hala 'sesli' olarak sınıflandıran hataları, toplam sessiz bölge örnek sayısı cinsinden, yüzdelik oranda ölçer (Eş. 10). Eş. 10'da, N_o, orjinal konuşmada ifade bitip sessiz bölgeye geçilirken, test edilen VAD metodu tarafından hala 'sesli' olarak tanımlanan giriş sinyal örneklerinin sayısı, $N_{silence}$, toplam sessiz bölge örnekleri sayısıdır.

$$OVER = \frac{N_O}{N_{silence}} x100 \quad (10)$$

3.5.3. Cümle ortası kırpma hatası (MSC-Middle sentence clipping)

Cümlelerin ortasındaki bölgeleri 'sessiz' bölge olarak yanlış sınıflandıran hataları, toplam sesli örnek sayısı cinsinden, yüzdelik oranda ölçer (Eş. 11). Eş. 11'de, N_M, orjinal konuşmada ifade devam ederken, test edilen VAD metodu tarafından ifade arasında sessiz bölge olarak tanımlanan giriş sinyal örneklerinin sayısı, N_{speech}, toplam konuşma örnekleri sayısıdır.

$$MSC = \frac{N_M}{N_{speech}} x100 \quad (11)$$

3.5.4. Gürültüyü konuşma olarak algılama (NDS-Noise detected as speech)

Sessiz bölgelerin ortasında 'sesli' olarak yanlış sınıflandıran hataları, toplam konuşma örnek sayısı cinsinden, yüzdelik oranda ölçer (Eş. 12). Eş. 12'de, N_N, orjinal konuşmada sessiz bölge devam ederken, test edilen VAD metodu tarafından 'sesli' olarak tanımlanan giriş sinyal örneklerinin sayısı, $N_{silence}$, toplam sessiz bölge örnekleri sayısıdır.

$$NDS = \frac{N_N}{N_{silence}} x100 \quad (12)$$

Testler sonucunda, konuşma olan bölgelerdeki toplam kırpma (clipping) hatası oranı (FEC+MSC) ile, konuşma olmayan bölgelerdeki toplam ekleme (insertion) hatası oranı ise (OVER+NDS)

ile ölçülür. Şekil 5, bir VAD karar sonucu örneği üzerinde objektif parametrelerin (NDS, FEC, OVER, MSC) bölgelerini göstermektedir.

4. Deneysel Test Düzeneği (Experimental Test Setup)

Bu makalede incelenen VAD metotları, farklı SNR seviyelerinde gürültülü giriş sinyalleri ile test edilerek karşılaştırılmış ve performansları ölçülmüştür. Tüm metotlar, çeşitli türde akustik arka plan gürültülere sahip sesler üzerinde test edilmiştir. Karşılaştırmada kullanılan VAD1 detektörü, değerlendirme için hem enerji hesaplamasını hem de ZCR'yi kullanmakta ve sabit bir eşik değer üzerinden karar üretmektedir [37]. VAD2 detektörü, VAD kararı verirken sadece enerji değerini kullanmaktadır ve bir giriş sinyalindeki değişen gürültü koşullarına karşı uyarlanabilir eşik değer avantajına sahiptir [48]. G.729B VAD detektörü [32, 45, 46], VAD1 ve VAD2'den farklı olarak, diferansiyel güç hesapları ve çizgi spektrum frekanslarını da (LSF) hesaba katarak belirlenen bir eşik değere göre VAD kararı verilmektedir. VAD3 detektörü ise, hem uzun vadeli spektral zarf ve ortalama gürültü spektrumunu karşılaştırılarak uzun bir analiz penceresi kullanmakta hem de eşik değeri ölçülen gürültü enerjisine adapte ederek karar üretmektedir [2].

Deneysel çalışmalarda, algoritmaları yazmak ve test etmek için MATLAB platformu kullanılmıştır. Algoritmanın performansı, doğru algılama yüzdesi, öznel konuşma kalitesi, arka plan gürültüsüne karşı sağlamlık temelinde analiz edilmiştir. Deneysel çalışmada kullanılan VAD metotları, üç erkek ve üç kadın konuşmacı tarafından telaffuz edilen 30 fonetik dengeli cümle içeren gürültüsüz ve gürültülü konuşma grupları (NOISEUS veri tabanı) [49, 50] ile değerlendirilmiştir. Veri tabanındaki gürültüsüz konuşma dosyaları, 4 farklı SNR'de (0dB, 5dB,10dB,15dB) AURORA veri tabanından yapay olarak eklenen yedi farklı ortam sesi (araba, restoran, gevezelik, havaalanı, sokak, tren istasyonu ve sergi sesleri) ile bozulmuştur. Veri tabanındaki sesler 8kHz'de örneklenmiştir. Deneysel çalışmada, söz konusu akustik gürültülü sesler bir araya getirilerek, 4 farklı SNR seviyesinde toplam yaklaşık 40 min. süreli test verisi sağlanmıştır. VAD algoritmalarını uygulamak ve test etmek için MATLAB sürümü (Vers. R2016a) kullanılmıştır. Referans olarak VAD kararı, VAD detektörünün temiz ses için en geniş aralıkta konuşma sinyalini yakalaması dikkate alınarak, elle işaretleme yapılmıştır. Detektörlerin VAD performansı, konuşma olan bölgeleri doğru algılamadaki ve sessiz bölgeyi doğru algılamadaki yüzdelik değerleri bakımından ölçülmüştür.

Testlerde, gürültüsüz konuşma sinyalinin konuşma olan bölgelerinin el ile işaretlenmesiyle elde edilen referans VAD kararları, 0dB, 5dB, 10dB ve 15dB SNR değerlerine sahip her çevresel gürültüdeki VAD'ların çıktılarıyla karşılaştırılmıştır. Sonuçlar her bir VAD algoritması için aşağıdaki bölümlerde gösterilmektedir. NOIZEUS veri tabanı ile yapılan testler ve sonuçları Tablo 1'de görülmektedir.

Gürültülü konuşma sinyalindeki SNR değişimine göre tespit performansının nasıl etkilendiği, konuşma olmayan bölgelerin doğru tespit oranı (HR0) (Eş. 13) ve konuşma bölgelerin doğru tespit oranı (HR1) (Eş. 13) olarak ölçülmüştür.

$$HR0 = \frac{N_0}{N_0^{ref}}, HR1 = \frac{N_1}{N_1^{ref}}, \quad (13)$$

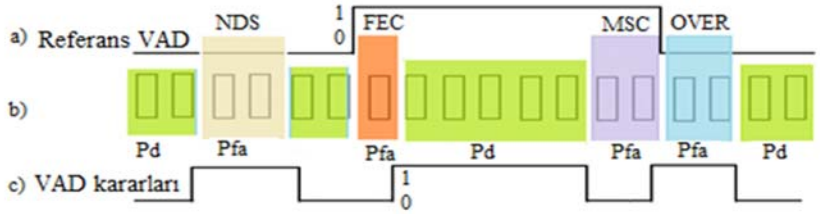
N_0^{ref} ve N_1^{ref}, referans alınan temiz konuşma sinyalinin toplam konuşma olan (VAD=1) ve konuşma olmayan (VAD=0) bölgelerini içermektedir. N0 ve N1 ise, değerlendirilen VAD metodunda tespit edilen konuşma olmayan ve konuşma olan bölgelerin sayısıdır.

Şekil 6 ve Şekil 7 sırasıyla, analiz edilen detektörlerin HR0 ve HR1 analiz sonuçlarını sunmaktadır. Farklı parametrelerin kıyaslamalı olarak detaylı analizi bakımından detektörlerin objektif performans ölçüm parametreleri Tablo 1'de sunulmuştur. Yapılan analizlerde, giriş sinyalindeki değişen akustik gürültü SNR seviyesine göre detektörlerin HR0 ve HR1 tespit yüzdelerindeki değişimlerin üzerinde odaklanılmıştır. Şekillerden ve Tablo 1'den çıkarılan sonuç aşağıdaki şekilde özetlenebilir.

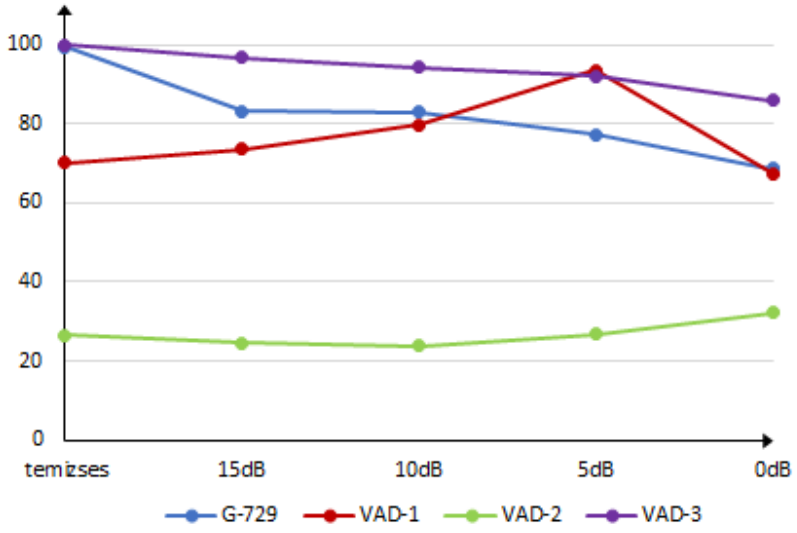
G.729B detektörü, temiz ses sinyali için HR1 bölgelerinin tespitinde %99 civarı tespit oranı ile oldukça başarılıdır. Ancak konuşma sinyalindeki gürültü oranı artıkça tespit oranı hızla düşmekte ve örneğin 0dB giriş sinyalleri için %70 civarına inmektedir. HR0 tespit oranındaki davranışı ise, HR1 tespit oranına göre daha düşük seyretmektedir.

VAD1 detektörünün temiz ses için %70 civarlarındaki HR1 tespit oranı örneğin 5dB SNR seviyesindeki giriş sesleri için %90 üzerine çıkmakta, 0dB giriş sesleri için tekrar %67 seviyelerine inmektedir. Yine bu seviyeler için artan SNR seviyelerine göre HR0 tespit oranındaki %50 civarı düşüş, VAD1 detektör performansının artan gürültüyle kötüleştiğinin, bu koşullarda HR1 değerlerindeki tutarsız değişim ise, detektörün artan gürültü durumunda VAD=1 kararı vermeye meyilli bir davranış sergilediğinin ve gürültülü koşullarda VAD bölgelerinin güvenilir şekilde tespit yeteneğini kaybettiğinin göstergesi olarak değerlendirilmektedir. ZCR ve enerji değerleri hesaplanması G.729B ve VAD1 algoritmalarında benzerdir. VAD1 algoritması, enerji ve ZCR değerinden bir VAD kararı üretmekte, sadece karar güçlüğü çekilen durumlarda analiz penceresini ikiye bölmektedir. Ancak, G.729B kodlayıcının alçak frekans bölgesi ve tüm frekans bölgesindeki iki ayrı diferansiyel güç hesabından bir karar üretmesi, VAD1'e göre gürültülü ortamlarda daha iyi performans sergilemesine yol açtığı değerlendirilmektedir.

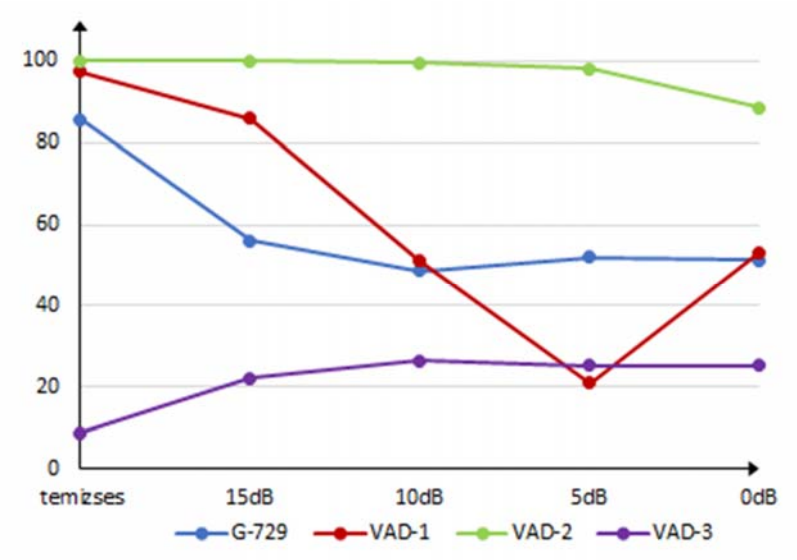
VAD2 detektörü, HR0 tespitinde gürültülü ortamlarda dahi çok başarılı performans sergilemektedir. HR1 tespitinde ise, tam olarak konuşma sinyalinin başlangıç-bitiş bölgelerini yakalamakta başarılı



Şekil 5. VAD tespit sonucunun NDS, FEC, OVER, MSC olarak değerlendirildiği bölgeler (Pfa: Yanlış Kabul olasılığı, Pd: doğru Kabul olasılığı) a) Referans VAD sinyali, b) Çerçevelenmiş sinyal kesitleri üzerinde NDS, FEC, OVER, MSC bölgelerinin gösterimi, c) Uygulanan test sonucundaki VAD kararları
(Regions where VAD detection result is evaluated as NDS, FEC, OVER, MSC (Pfa: False Acceptance probability, Pd: Correct Acceptance probability) a) Reference VAD signal, b) Display of NDS, FEC, OVER, MSC regions on framed signal sections, c) VAD decisions as a result of the applied test)



Şekil 6. Test edilen VAD'ların HR1 tespit oranları (HR1 hit rates for tested VADs)



Şekil 7. Test edilen VAD'ların HR0 tespit oranları (HR0 hit rates for tested VADs)

olmamakla birlikte, HR1 tespit oranının 0dB civarlarında dahi temiz-ses için sergilediğine benzer performans sergilemesinin, detektör yapısının tam olarak konuşma başlangıç-bitiş noktalarını yakalamak yerine konuşma olan bölgelerin tespitine odaklandığı ve bu konuda başarılı bir performans sergilediği şeklinde değerlendirilmektedir. Bu bakımdan, gürültü miktarı arttığı halde HR1 tespit oranındaki değişim oldukça yavaş ve kararlı görünmektedir. VAD2'nin HR1 tespitindeki Şekil 6'da görünen düşük oranın diğer bir nedeninin, testlerdeki referans alınan VAD sinyalinin konuşma sinyalini en geniş aralıkta yakalayabilmesi için eşik değere oldukça yakın seçilmesi olduğu değerlendirilmektedir. VAD2 detektörünün HR0 tespit oranındaki yüksek başarı oranı, algoritmanın gürültülü ortamlarda konuşma olmayan bölgeleri yanlışlıkla konuşma olarak etiketlemesindeki oranın düşüklüğünü göstermektedir ve bu bakımdan diğer detektörlere göre oldukça başarılıdır.

VAD3 detektörü, HR1 tespit oranı bakımından, temiz ses için %99 civarı tespit oranının 0dB civarlarında %85'e inmesi ile tüm gürültülü koşullardaki giriş sinyallerinde HR1 tespit oranında oldukça başarılı performans sergilemekte, ancak %25 civarı tespit değeri ile, HR0 tespit oranındaki başarısının düşük olduğu gözlemlenmektedir. VAD3 detektörünün, uzun pencere aralığındaki LTSD analizinin HR1 oranını arttırmada başarılı olduğu değerlendirilmektedir. VAD3 detektörü için HR0 ve HRI yüzdelik miktarlarındaki [2]'ye göre oluşan farklılıkların, hem test olarak seçilen veri kümelerinin farklılığından hem de referans olarak kullanılan VAD sinyallerinin her iki çalışmada da elle işaretlenmesinden kaynaklanan farklılıklar olabileceği değerlendirilmektedir. Ancak hem G.729B hem de VAD3 detektörlerinin değişen SNR koşullarına davranış karakteristiklerinin her iki çalışmada da benzer olması, bu çalışmada yapılan değerlendirmenin tutarlılığını ortaya koymada bir gösterge olarak değerlendirilmektedir.

5. Sonuçlar ve Tartışmalar (Results and Discussions)

Bu çalışmada, VAD algoritmalarının farklı akustik gürültülü ortamlardaki performanslarını etkileyen sebepleri ortaya koyabilmek için literatürdeki bazı konuşma aktivite detektörleri (VAD)

Tablo 1. NOISEUS veri tabanı için, VAD metotlarının, FEC, MSC, OVER, NDS hata ve HR0, HR1 doğru tespit oranları
(For NOISEUS database, FEC, MSC, OVER, NDS error and HR0, HR1 correct detection rates of VAD methods)

Uygulanan Metot	Giriş Konuşma sinyali	FEC (Ön uç kırpma oranı)	MSC (Cümle kırpma oranı)	OVER (Taşma hatası oranı)	NDS (Gürültü tespit hata oranı)	(FEC+MSC) (Toplam kırpma hatası oranı)	(OVER+NDS) (Toplam ekleme hatası oranı)	HR1 (Konuşma bölgesi tespit oranı)	HR0 (Sessiz bölge tespit oranı)
G.729B	Temizses	0,499	0,029	14,374	0,000	0,528	14,374	99,472	85,626
	15dB	4,189	12,524	31,115	12,881	16,713	43,996	83,287	56,004
	10dB	4,638	12,305	37,859	13,501	16,943	51,360	83,057	48,640
	5dB	5,885	16,771	33,899	14,159	22,656	48,058	77,344	51,942
	0dB	8,507	22,764	36,189	12,514	31,271	48,703	68,729	51,297
VAD1	Temizses	9,775	20,218	0,531	1,949	29,993	2,480	70,007	97,520
	15dB	7,740	18,791	6,441	7,681	26,531	14,121	73,469	85,879
	10dB	5,714	14,666	35,607	13,198	20,380	48,804	79,620	51,196
	5dB	1,760	4,817	70,555	8,402	6,577	78,957	93,423	21,043
	0dB	10,811	21,822	35,746	11,110	32,634	46,856	67,366	53,144
VAD2	Temizses	23,287	50,284	0,000	0,000	73,571	0,000	26,429	100,000
	15dB	24,376	51,167	0,000	0,000	75,543	0,000	24,457	100,000
	10dB	25,374	50,774	0,013	0,316	76,148	0,329	23,852	99,671
	5dB	23,200	50,022	0,089	1,860	73,222	1,949	26,778	98,051
	0dB	19,239	48,605	4,125	7,288	67,844	11,413	32,156	88,587
VAD3	Temizses	0,000	0,073	80,514	10,629	0,073	91,143	99,927	8,857
	15dB	0,842	2,392	61,584	16,298	3,233	77,882	96,767	22,118
	10dB	1,575	4,148	59,636	13,868	5,723	73,504	94,277	26,496
	5dB	2,242	5,743	62,533	12,261	7,985	74,794	92,015	25,206
	0dB	4,790	9,249	62,875	11,793	14,038	74,668	85,962	25,332

incelenmiştir. Makaledeki VAD detektörleri, test veri tabanında yer alan farklı SNR seviyelerindeki akustik gürültülü koşullar altında test edilmiştir. Detektörlerin VAD performansı, ilgili detektörün karar aşamasında kullanılan eşik değerinin sabit veya uyarlamalı olması, seçilen analiz penceresinin kısa veya uzun olması, birden fazla özellik vektörünün birlikte kullanımı gibi durumlarının sonuçtaki VAD kararına etkisi bakımından değerlendirilmiştir. Bu kapsamda, gürültüsüz bir konuşma sinyali referans alınarak konuşma tespit oranı (HR1) veya sessizlik tespit oranı (HR0) tespit bölgeleri el ile etiketlenmiştir ve VAD detektörlerinin HR0 ve HR1 tespit bölgeleri ile karşılaştırılmıştır. Tablo 1'deki sonuçlar incelendiğinde kısaca şu tespitler yapılmıştır; VAD1 detektörü, HR1 tespit oranında gürültüye karşı performansı incelendiğinde, kararsız bir seyir izlemektedir. Temiz ve düşük gürültülü giriş sinyalleri için G.729B detektörü, HR1 ve HR0 tespitinde başarılı bir performans sergilemekle birlikte, gürültü düzeyi arttıkça HR1 ve HR0 tespit oranları hızla düşmektedir. Diğer yandan G.729B detektörü gürültüye bağlı performans davranışında VAD1'e göre daha tutarlı bir seyir izlenmektedir. G.729B ve VAD1 detektörlerinin VAD kararında kullandığı enerji ve ZCR hesaplamaları benzer olmasına rağmen, sonuç performanstaki farklılığın G.729B detektörünün VAD kararında kullandığı diğer bir özellik vektörü olan çizgi spektrum frekanslarından (LSF) kaynaklandığı değerlendirilmektedir.

VAD2 detektörünün HR1 değerleri incelendiğinde, diğer detektörlere göre daha düşük değerde görünmektedir. Ancak, tüm gürültü seviyelerinde HR1 değerindeki değişim oranının düşüklüğü, VAD2 detektörünün test edilen tüm akustik gürültülü koşullarda konuşma bölgelerinin tespit edilmesindeki performansını koruduğunu göstermektedir. VAD2 detektörü, konuşma sinyalinin başlangıç-bitiş bölgelerini tam olarak yakalamakta başarılı olmamakla birlikte, konuşma olan bölgelerin tespitinde tutarlı ve gürültüye dayanıklı bir performans sergilemektedir. Ayrıca VAD2 detektörü, HR0 tespitindeki gürültüye karşı sergilediği dayanıklılıkla, tüm akustik gürültülü ortamlarda konuşma olmayan bölgelerin yanlışlıkla konuşma olarak etiketlemesinde hususunda diğer metotlara göre çok başarılı performans sergilediği değerlendirilmektedir. VAD2 detektörünün gürültüye karşı üstün performansının, büyük oranda uyarlamalı eşik değer hesabından kaynaklandığı değerlendirilmektedir.

VAD3 detektörü, HR1 tespit oranı ve akustik gürültü dayanıklılığı bakımından incelendiğinde hem yüksek tespit oranı hem de gürültüye karşı dayanıklılık bakımından diğer test edilen detektörlere göre üstün bir performans ortaya koymuştur. Diğer yandan, HR0 tespit oranındaki başarısının ise düşük olduğu gözlenmiştir. VAD3 detektörünün, uzun pencere aralığındaki LTSD analizinin HR1 oranını arttırmada başarılı olduğu değerlendirilmektedir. VAD3 detektörü, VAD2 detektörü gibi uyarlamalı bir eşik değer hesabı kullanmaktadır. Buna rağmen, VAD3 detektörünün VAD2'ye göre HR0 oranındaki daha düşük değerlerde olması, VAD3 detektörünün (VAD=1) kararı vermeye meyilli davranış sergilediği olarak değerlendirilmektedir.

Bu makaledeki deneysel çalışmada kullanılan VAD metotları gürültü dayanıklılık performansları bakımından değerlendirilmiş ve bu kapsamda geniş bir akustik gürültülü konuşma veri kümesi içermesi bakımından NOISEUS veri tabanı [49, 50] kullanılmıştır. İleri bir çalışma olarak, VAD metotlarının özel bazı konuşma veri kümelerine karşı performanslarının ölçülmesi için bu kapsamdaki veri kümelerinden faydalanılabilir. Son yıllarda ülkemizde hazırlanarak konuşma tanıma sistemlerinde kullanılan konuşma veri kümelerinden bazıları şu şekildedir; Türkçe ağızları veri kümesi kullanılarak büyük ölçekli konuşma tanıma sistemlerinin başarısını arttırmak için, uzun-kısa dönem bellekli (LSTM) sinir ağları kullanılarak ağız tanıma yapılmış ve yüksek başarı oranı elde edilmiştir [51]. Diğer bir konuşma veri kümesi, işitme engelli kişilere yönelik geniş bir veri kümesi derlenmiş ve bu veri kümesi ile sınıflandırma modeli oluşturulmuştur. Sınıflandırma modeli için evrişimli sinir ağı (CNN) modeli kullanılmıştır [52]. Türkiye'nin çeşitli bölgelerinde konuşulan

ağızlara yönelik olarak veya konuşma güçlüğü içindeki insanlara yönelik olarak hazırlanan bu özel konuşma veri kümelerinin VAD sistemlerinde giriş sinyali olarak kullanılması, sistemlerin etkinliği hakkında yeni veriler ortaya koyabilecektir.

Bu makalede, farklı karakteristik özelliklerdeki VAD detektörleri ele alınarak, değişen akustik gürültülü koşullarda performanslarına etki eden faktörler test edilmiştir. Bu kapsamda, gürültülü akustik giriş sinyallerindeki SNR değeri değiştikçe, VAD detektörlerinin tespit özelliklerindeki farklılaşmanın sebepleri araştırılmıştır. Ele alınan VAD detektörlerinin performans değerlendirmesinde, özellik vektörleri hesaplama metotlarının karar aşamasına etkileri, birden fazla özellik vektörünün birlikte kullanımının sonuç performansa etkileri, uyarlamalı veya sabit eşik değer kullanılmasının karar sonucu üzerindeki etkileri, analiz penceresinin kısa veya uzun seçilmesinin etkileri gibi hususlar karşılaştırmalı olarak analiz edilmiştir. Deneysel çalışmadaki sonuçların farklı literatür verileriyle kıyaslamalarını kolaylaştırmak amacıyla, testler sırasında bu amaca yönelik olarak geliştirilmiş olan ve farklı SNR seviyelerinde çeşitli akustik gürültülü sesler içeren NOIZEUS veri tabanı kullanılmıştır. Testler objektif test ölçüm metotları kullanılarak yapılmış ve her bir VAD metodunun HR1 ve HR0 tespit doğruluğu ölçülmüştür. Tablo 1'de görüleceği üzere, objektif test sonuçlarının detaylı analiz edilebilmesi için HR1 ve HR0 oransal değerleri yanında, 4 farklı nesnel ölçüm parametre sonuçları da hesaplanarak sunulmuştur.

Bu çalışmadaki test sonuçları göstermiştir ki, her bir VAD detektörü değişen akustik gürültülü koşullarda hem HR1 hem de HR0 tespit oranı bakımından farklı dayanıklılık performansları sergilemiş, tüm akustik koşullarda tek başına yüksek performans sunan bir VAD detektörü tespit edilememiştir. Diğer yandan, VAD detektörlerinde uygulanan metotların değişik akustik gürültülü koşullardaki etkinliklerini görmek bakımından, deneysel sonuçlar oldukça dikkate değer bilgiler sunmuştur. Söz konusu bilgilerin, farklı akustik gürültülü koşullarda optimum performans sergileyebilen bir VAD algoritması tasarımında yol gösterici olacağı değerlendirilmektedir.

Kaynaklar (References)

1. Ramírez J. , Gorriz J. M. , Segura J. C., Voice Activity Detection. Fundamentals and Speech Recognition System Robustness In book: Robust Speech Recognition and Understanding, Edited by Michael Grimm, I-Tech Education and Publishing, June 2007.
2. Ramírez J., Segura J. C., Benítez C., Torre Á., Rubio A., Efficient voice activity detection algorithms using long-term speech information, Speech Communication, 42 (3–4), 271-287, 2004.
3. Benyassine A., Shlomot E., and Su H.-Y., ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application, IEEE Communication Magazine, 35, 64–73, Sept. 1997.
4. Junqua J. C., Reaves B., and Mark B., A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize, Proceedings of Eurospeech, 1371-1374, 1991.
5. Lamel F., Rabiner R., Rosenberg E., Wilpon G., An improved endpoint detector for isolated word recognition, IEEE Transactions on Acoust. Speech Signal Processing, 29, 777-785, 1981.
6. Kotnik B., Kacic Z., Horvat B.: A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm, Proceedings of 7th Europseech, 197-200, 2001.
7. Nakatani T., Irino T., and Zolfaghari P., Dominance spectrum based V/UV classification and F estimation, Proceedings of Eurospeech '03, 2003.
8. Ahmadi S. and Spanias A. S., Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Transactions on Speech Audio Process., 7 (3), 333–338, May 1999.
9. Chengalvarayan R., Robust energy normalization using speech/non-speech discriminator for German connected digit recognition, Proc. EUROSPEECH 1999, Budapest, Hungary, pp. 61–64., 1999.
10. Wu B.F., Wang K.C., Robust endpoint detection algorithm based on the adaptive band partitioning spectral entropy in adverse environments. IEEE Transactions on Speech Audio Processing 13, 762–775, 2005.
11. Nemer, E., Goubran, R., Mahmoud, S., Robust voice activity detection using higherorder statistics in the lpc residual domain, IEEE Transactions on Speech and Audio Processing, 9 (3), 217–231, 2001.
12. Kingsbury B., Saon G., Mangu L., Padmanabhan M., Sarikaya R., Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system, Proc. ICASSP, 53-56, 2002.
13. Kristjansson T., Deligne S., Olsen P., Voicing features for robust speech detection, Proceedings of Interspeech, 369-372, 2005.
14. Marzinzik M.; Kollmeier B., Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, IEEE Transactions on Speech and Audio Processing, vol. 10, no. 6, pp. 341–351, 2002.
15. McClellan S. A., Gibson J. D., Spectral entropy: An alternative indicator for rate allocation, IEEE International Conference on Acoustics, Speech, Signal Processing, (Adelaide, Australia), 201-204, Apr. 1994.
16. Kristjansson T., Deligne S., Olsen P., Voicing features for robust speech detection (INTERSPEECH, 369–372, 2005.
17. Noll A. M., Cepstrum pitch determination, Journal of the Acoustic Society of America, 41, 293-309, Feb. 1967.
18. Haigh J. A., Mason J. S., Robust voice activity detection using cepstral features, Proc. of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation, Beijing, China, 1993.
19. Chung K., Oh S. Y., Voice Activity Detection Using an Improved Unvoiced Feature Normalization Process in Noisy Environments, Wireless Personal Communications, 89, 3, 1-13, 2015.
20. Ahmadi S., Spanias A.S., Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Transactions on Speech Audio Processing 7, 333–338, 1999.
21. Stegmann J., Schroder, G.., Robust voice-activity detection based on the wavelet transform, in Speech Coding for Telecommunications Proceeding, 1997 IEEE Workshop on, 1997, 99-100., 1997.
22. Chen S. H., Wu H. T., Chang Y., Truong T. K., Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator, Pattern Recognition Letters, 28 (11), 1327-1332, 2007.
23. Sadjadi S. O., Hansen J. H. L., Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux, in IEEE Signal Processing Letters, 20 (3), 197-200, March 2013.
24. Tahmasbi R., Rezaei S., Change point detection in GARCH models for voice activity detection, IEEE Trans. Audio, Speech, Lang. Process., 16, 5, 1038–1046, Jul. 2008.
25. Davis A., Nordholm S., Togneri R., Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold, IEEE Transactions on Audio, Speech, Language Processing, 14 (2), 412–424, Mar. 2006.
26. Chang J., Kim N. K., Mitra S. K., Voice activity detection based on multiple statistical models, IEEE Transactions on Signal Processing, 54 (6), 1965-1976, Jun. 2006.
27. Haigh J., Mason J., A voice activity detector based on cepstral analysis, Proceedings of Eurospeech, pp. 1103-1106, 2003.
28. Eyben F., Weninger F., Squartini S., Schuller B., Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies, in Proc. ICASSP, 483-487, 2013.
29. Ferroni G., Bonfigli R., Principi E., Squartini S., Piazza F., A Deep Neural Network approach for Voice Activity Detection in multi-room domestic scenarios, International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015.
30. Bie F., Z., D. Wang, T. F. Zheng, DNN-based Voice Activity Detection for Speaker Recognition, CLST Technical Report, 1-11, 2015.
31. Ali Z., Talha M., Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments, IEEE Access, 6, 15494-15504, 2018.
32. Beritelli F., Casale S., Ruggeri G. and Serrano S., Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors, IEEE Signal Processing Letters, 9 (3), 85-88, March 2002.
33. GSM 06.94. Digital cellular telecommunication system (Phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech

traffic channels; general description. ETSI, Tech. Report, V.7.0.0, 1999.
34. Morales J. M., Design and Implementation on DSP of the ETSI GSM Adaptive Multi-Rate Vocoder, MsC thesis, the Faculty of Telecommunication Engineering, Universitat Politècnica de Catalunya (UPC), 2008.
35. Yoo H. L. and Yook D., Formant-Based Robust Voice Activity Detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23 (12), 2238-2245, Dec. 2015.
36. Rouat J., Liu Y. C., and Morrisette D., A pitch determination and voiced/unvoiced decision algorithm for noisy speech, Speech Communication, 21, 1997.
37. Bachu R. G., Kopparthi S., Adapa B., Barkana B. D., Voiced/Unvoiced Decision for Speech Signals Based on ZeroCrossing Rate and Energy, Advanced Techniques in Computing Sciences and Software Engineering, 279-282, January, 2010.
38. Brokish C. W., Lewis M., A-Law and μ-Law Companding Implementations Using the TMS320C54x, Texas instruments, 1997
39. Muralishankar R., Ghosh D. and Gurugopinath S., A Novel Modified Mel-DCT Filter Bank Structure With Application to Voice Activity Detection, in IEEE Signal Processing Letters, 27, 1240-1244, 2020.
40. Haigh J. A., Mason J. S., Robust voice activity detection using cepstral features, Proc. IEEE TENCON,'93. Proceedings of Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on, China, 3, 321-324, 1993.
41. Sehgal A., Kehtarnavaz N., A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. IEEE Access, (6) 9017-9026, 2018.
42. Ivry A., Berdugo B. and Cohen I., Voice Activity Detection for Transient Noisy Environment Based on Diffusion Nets, IEEE Journal of Selected Topics in Signal Processing, 13 (2), 254-264, May 2019.
43. Ali Z., Talha M., Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments, IEEE Access, 6, 15494-15504, 2018.
44. Lee Y., Min J., Han D. K., Ko H., Spectro-Temporal Attention-Based Voice Activity Detection, IEEE Signal Processing Letters, 27, 131-135, 2020.
45. Benyassine A., Shlomot E., Su H.Y., Yuen E., A robust low complexity voice activity detection algorithm for speech communication systems, 1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. PA, USA, 97-98, 1997.
46. Benyassine A., Shlomot E., and Su H.-Y., ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application, IEEE Commun. Mag., 35, 64–73, Sept. 1997.
47. Beritelli F., Casale S., Cavallaro A., A robust voice activity detector for wireless communications using soft computing, IEEE Journal on Selected Areas in Communications, 16, 1818–1829, Dec. 1998.
48. Sakhnov K., Verteletskaya E., Simak B., Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications, Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, London, 2009.
49. Loizou P., NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms, Speech Communication, 49, 588-601, 2017.
50. Hirsch H.G., Pearce D., The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, ISCA Tutorial and Research Workshop (ITRW) ASR2000, September 2000.
51. Işık G., Artuner H., Turkish dialect recognition in terms of prosodic by long short-term memory neural networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (1), 213-224, 2020.
52. Özcan T., Baştürk A., ERUSLR: A new Turkish sign language dataset and its recognition using hyperparameter optimization aided convolutional neural network, Journal of the Faculty of Engineering and Architecture of Gazi University, 36 (1), 527-542, 2021.