Yapay Zeka Tabanlı Ses Sistemlerinde Gürültü Engelleme (AI-Based Noise Cancelling)

Mert Şahin
Bilgisayar Mühendisliği
Ankara Üniversitesi
Ankara, Türkiye
mert096488@gmail.com

***Özetçe*—Bu proje, Yapay Zeka Tabanlı Ses Sistemlerinde Gürültü Engelleme (AI-Based Noise Cancelling) konusuna odaklanmaktadır. Geliştirilecek olan yazılım sayesinde arka plandaki istenmeyen gürültüler giderilerek operasyon sırasında iletişimin kalitesinin artırılması hedeflenmektedir. Projede, ses iyileştirme için difüzyon olasılıksal modellerinde koşul bilgisini dahil etme zorluğunu ele alan DOSE modeli kullanılacaktır ve Türkçe konuşmalardan oluşan bir veri seti üzerinde çalışacaktır.**

Anahtar Kelimeler— Yapay Zekâ, Gürültü Engelleme, Ses İyileştirme, DOSE

I. GİRİŞ

Bu proje, yapay zekâ tabanlı konuşma geliştirme (SE) yöntemleri üzerine odaklanarak, kulaklık, mikrofon gibi ses sitemlerinde değişen koşullara adaptif, yüksek kaliteli iletişim sağlamayı hedeflemektedir. Günümüzde, sesle kontrol edilen cihazlar, konuşmadan metne dönüştürme, telekonferans sistemleri, uzaktan eğitim platformları, müşteri hizmetleri çağır merkezleri ve hatta sürücüsüz araçlardaki sesli komut sistemleri gibi birçok alanda konuşma anlaşılırlığı kritik öneme sahipken, arka plan gürültüsü bu sistemlerin performansını ciddi şekilde düşürmektedir. Karşılaşılan yaygın gürültü türleri arasında Gaussian gürültüsü, darbe gürültüsü, çevresel gürültü, yapısal gürültü ve kanal kaynaklı gürültüler bulunmaktadır[1]. Bu çeşitlilik, tek tip bir gürültü azaltma yaklaşmanın yetersiz kalmasına neden olmaktadır.

Geleneksel konuşma geliştirme yaklaşımlarında deterministik derin ögrenme modelleri kullanılsa da , son araştırmalar gürültü azaltıcı difüzyon olasılıksal modelleri (DDPM'ler) gibi üretken yaklaşımların da oldukça etkili olabileceğini göstermiştir[2,3]. Ancak DDPM'lere koşul bilgisinin (yani temiz konuşma sinyalinin özelliklerini veya gürültünün türünü) dahil etmek, SE alanında önemli bir zorluk teşkil etmektedir[4,5,6]. Bu zorluğun üstesinde gelmek için, modele özgü olmayan ve iki verimli koşul artırma tekniğini kullanan DOSE adında bir yöntem kullanılacaktır[7]. Bu yöntem, modelin eğitim aşamasında dropout (açılma) kullanarak örnekler üretirken koşul faktörüne öncelik vermesini sağlamakta ve bilgilendirici, uyarlanabilir bir öncül sağlayarak koşul bilgini örnekleme sürecine daha etkin bir şekilde dahil etmektedir.

Bu projenin en önemli yeniliği, mevcut derin ögrenme modellerinin genellikle Türkçe kaynaklarla eğitilmemesi problemine çözüm getirmesidir. Geliştirilecek yöntemler, özel olarak toplanacak ve etiketlenecek bir Türkçe ses veri kümesi üzerinden eğitilerek, gürültülü ortamda Türkçe konuşmayı daha net ve anlaşılır hale getirmeyi hedeflemektedir.

II. LİTERATÜR TARAMASI

Projemiz, temelde iki ana projeden ilham almıştır. Bu projeler, kullandıkları model mimarileri doğrultusunda birbirlerinden ayrışmaktadırlar.

A. *Deep Learning-Based Speech Enhancement for Robust Speech Recognition in Noisy Enviroments(Sowmya vd., 2025)[8]*

Bu çalışma, gürültülü ortamda ASR performansını arttırmak için derin ögrenme tabanlı konuşma iyileştirmeyi amaçlamaktadır. Makale, "Mud Bas" adli, MATLAB ile eğitilmiş bir CNN tabanlı DNN modeli önermektedir. Bu model, gürültülü konuşmayı temizleyerek, tam olarak hangi veri seti ile çalıştığını belirtmese de , farklı gürültü türleri ve seviyeleri içeren geniş bir gürültü ve temiz konuşma veri seti üzerinde yapılan deneylerde, SNR, PESQ ve MOS gibi metriklerde geleneksel yöntemlerden daha iyi performans göstererek derin ögrenmenin gürültü ortamlara genelleme yeteneğini kanıtlamaktadır. Fig1 : Diğer modellere göre gürültü azaltma performansını gösteriyor.

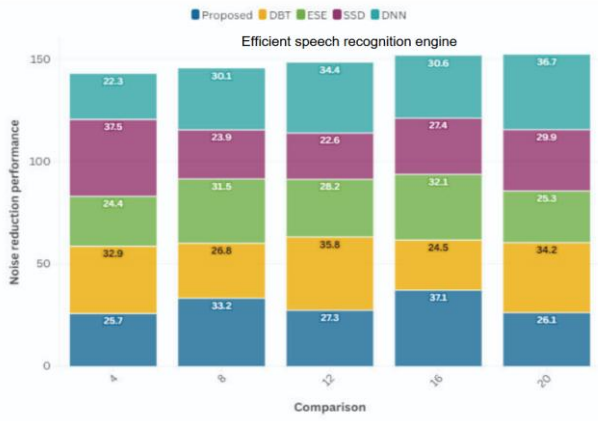

Fig. 1 Gürültü Azaltama Performansının Hesaplanması [8]

B. *DOSE: Diffusion Dropout with ADaptive Prior for Speech Enhancement (Xinchi Zhang vd., 2023)[7]*

Bu çalışma, difüzyon modellerinin konuşma iyileştirmedeki genelleme sorununu ve aşırı temizlenmeyi çözmeyi amaçlamaktadır. Bu makale difüzyon modellerinde "dropout" ve uyarlanabilir önseli birleştiren DOSE çerçevesini önermektedir. Figüre 2 de önerilen DOSE yöntemleri görülmektedir.

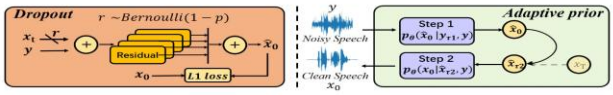

Fig. 2 Önerilen DOSE Yöntemleri [7]

Voicebank-DEMAND[11] ve WSJO-CHIME4 gibi çeşitli veri setleri üzerinde test edilen DOSE, objektif ve subjektif metriklerde mevcut modellerden üstün performans göstererek daha doğal ve iyi genellenebilen iyileştirmiş konuşma üretme yeteneğini sergilemektedir.

Method	Year	Efficiency	Dataset	STOI(%)↑	PESQ↑	CSIG↑	CBAK↑	COVL↑
Unprocessed	–	–	VB	92.1	1.97	3.35	2.44	2.63
DiffWave	2021	1 step (dis)		93.3	2.51	3.72	3.27	3.11
DiffuSE	2021	6 steps		$93.5^{+0.20}_{\pm0.05}$	$2.39^{-0.12}_{\pm0.01}$	$3.71^{-0.01}_{\pm0.01}$	$3.04^{-0.23}_{\pm0.01}$	$3.03^{+0.08}_{\pm0.01}$
CDiffuSE	2022	6 steps		$93.7^{+0.40}_{\pm0.05}$	$2.43^{-0.08}_{\pm0.01}$	$3.77^{+0.05}_{\pm0.01}$	$3.09^{-0.18}_{\pm0.01}$	$3.09^{-0.02}_{\pm0.01}$
SGMSE	2022	50 steps		$93.3^{+0.00}_{\pm0.08}$	$2.34^{-0.17}_{\pm0.01}$	$3.69^{-0.03}_{\pm0.01}$	$2.90^{-0.37}_{\pm0.01}$	$3.00^{-0.11}_{\pm0.01}$
DR-DiffuSE	2023	6 steps		$92.9^{-0.04}_{\pm0.06}$	$2.50^{+0.01}_{\pm0.02}$	$3.68^{-0.04}_{\pm0.02}$	$3.27^{+0.00}_{\pm0.02}$	$3.08^{-0.03}_{\pm0.02}$
DOSE	–	2 steps		$93.6^{+0.30}_{\pm0.05}$	$2.56^{+0.05}_{\pm0.01}$	$3.83^{+0.11}_{\pm0.01}$	$3.27^{+0.00}_{\pm0.01}$	$3.19^{+0.08}_{\pm0.01}$
Unprocessed	–	–	CHIME-4	71.5	1.21	2.18	1.97	1.62
DiffWave	2021	1 step (dis)		72.3	1.22	2.21	1.95	1.63
DiffuSE	2021	6 steps		$83.7^{+11.4}_{\pm0.05}$	$1.59^{+0.36}_{\pm0.01}$	$2.91^{+0.70}_{\pm0.01}$	$2.19^{+0.24}_{\pm0.01}$	$2.19^{+0.56}_{\pm0.01}$
CDiffuSE	2022	6 steps		$82.8^{+10.5}_{\pm0.05}$	$1.58^{+0.36}_{\pm0.01}$	$2.88^{+0.67}_{\pm0.01}$	$2.15^{+0.20}_{\pm0.01}$	$2.18^{+0.55}_{\pm0.01}$
SGMSE	2022	50 steps		$84.5^{+12.2}_{\pm0.05}$	$1.57^{+0.34}_{\pm0.02}$	$2.92^{+0.71}_{\pm0.01}$	$2.18^{+0.23}_{\pm0.02}$	$2.18^{+0.55}_{\pm0.01}$
DR-DiffuSE	2023	6 steps		$77.6^{+6.30}_{\pm0.06}$	$1.29^{+0.07}_{\pm0.04}$	$2.40^{+0.19}_{\pm0.02}$	$2.04^{+0.09}_{\pm0.01}$	$1.78^{+0.15}_{\pm0.01}$
DOSE	–	2 steps		$86.6^{+14.3}_{\pm0.05}$	$1.52^{+0.30}_{\pm0.01}$	$2.71^{+0.58}_{\pm0.01}$	$2.15^{+0.20}_{\pm0.01}$	$2.06^{+0.43}_{\pm0.01}$

Tablo 1 : Farklı difüzyon iyileştirme yöntemlerinin karşılaştırması [7]

III. METODOLOJİ

A. *Veri Seti Toplama ve Ön İşleme*

Projenin başarısı için kaliteli ve çeşitli veri setleri kritik öneme sahiptir. Bu bağlamda, özellikle Türkçe konuşmaları içeren temiz konuşma verisi olarak Hugging Face üzerinden temin edilecek issai/Turkish_Speech_Corpus[13] veri seti kullanılacaktır. Ardından, gürültü sesleri için Microsoft Scalable Noisy Speech Dataset (MS-SNSD)[14] veri setinden faydalanılarak, bu temiz Türkçe konuşma verileri ile farklı gürültü türleri ve seviyeleri (Sinyal-Gürültü Oranı - SNR) içeren sentetik olarak gürültülendirilmiş yeni bir veri seti oluşturulacaktır. Bu yeni veri seti, modelin eğitimi ve test edilmesi için kullanılacaktır.

B. *DOSE Çerçevesinin Uygulanması ve Uyarlanması*

Projenin temelini oluşturan ve GitHub üzerinden erişilen modele ve kod tabanına[12] dayandırılacak olan DOSE çerçevesinin prensipleri (Diffusion Dropout ve Adaptive Prior mekanizmları) üzerinde yoğunlaşılacaktır[7]. Gürültü konuşmanın koşullu dağılımını daha iyi ögrenmek ve aşırı temizlenmeyi önlemek amacıyla, bu prensiplerin proje hedeflerine uygun şekilde uygulanması ve optimize edilmesi hedeflenmektedir.

C. *Model Eğitimi ve Optimizasyonu*

Tasarlanan difüzyon modeli, önceden hazırlanan gürültülü ve temiz Türkçe konuşma veri setleri üzerinden eğitilecektir. Eğitim süreci boyunca, dropout işlemi için dropout olasılığı parametresi ve uyarlamalı öncül (adaptive prior) için iki zaman adımı t1 ve t2 parametreleri, DOSE makalesindeki öneriler ışığında ayarlanacak ve modelin performansı optimize edilecektir.

D. *Değerlendirme ve Analiz*

Eğitilen modelin performansı, literatürde yaygın olarak kullanılan objektif ölçütler olan Segmental Sinyal-Gürültü Oranı (SSNR), Konuşma Kalitesinin Algısal Değerlendirmesi (PESQ)[9], Kısa Sureli Nesnel Anlaşılabilirlik İndeksi (STOI)[10] ve Ortalama Görüş Puanı (MOS) gibi kullanılarak değerlendirilecektir. Özellikle, DOSE makalesinde elde edilen sonuçlarla karşılaştırmalı analizler yapılacaktır.

IV. SONUÇ

Projemiz, gürültülü ortamlara ASR sistemlerinin ihtiyacından doğmuştur. Temelini DOSE çerçevesinden alan bu çalışma, literatüre doğrudan katkı yerine, DOSE modelini Türkçe konuşma verileri (issai/Turkish_Speech_Corpus) ve MS-SNSD gürültü veri setleri kullanılarak oluşturulan sentetik bir veri seti üzerinden uygulayıp analiz etmeyi hedeflemektedir. Bu sayede, modelin Türkçe bağlamdaki performansı gösterilerek dil odaklı konuşma iyileştirme uygulamalarına pratik bir örnek sunulacaktır.

REFERENCES

[1] Singasani, T. R., Shaik, M. S., Gangajaliya, C., Ogety, S. S., Nallam, V. A. B., & Katta, S. K. R. (2025, June). AI-Driven Signal Processing: Improving Communication Systems with Machine Learning-Based Noise Reduction. In 2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN) (pp. 1-6). IEEE.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851.

[3] Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761.

[4] Tai, W., Zhou, F., Trajcevski, G., & Zhong, T. (2023, June). Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13627-13635)

[5] Lu, Y. J., Wang, Z. Q., Watanabe, S., Richard, A., Yu, C., & Tsao, Y. (2022, May). Conditional diffusion probabilistic model for speech enhancement. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7402-7406). Ieee.

[6] Serrà, J., Pascual, S., Pons, J., Araz, R. O., & Scaini, D. (2022). Universal speech enhancement with score-based diffusion. arXiv preprint arXiv:2206.03065.

[7] Tai, W., Lei, Y., Zhou, F., Trajcevski, G., & Zhong, T. (2023). DOSE: Diffusion dropout with adaptive prior for speech enhancement. Advances in Neural Information Processing Systems, 36, 40272-40293.

[8] Sowmya, C. S., Das, N., Sharma, D., Mondal, S., Soni, I., & Kumar, N. (2025, March). Deep Learning-Based Speech Enhancement for Robust Speech Recognition in Noisy Environments. In 2025 International Conference on Automation and Computation (AUTOCOM) (pp. 1076-1081). IEEE.

[9] Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001, May). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221) (Vol. 2, pp. 749-752). IEEE.

[10] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing (pp. 4214-4217). IEEE.

[11] Veaux, C., Yamagishi, J., & King, S. (2013, November). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE) (pp. 1-4). IEEE.

[12] https://github.com/ICDM-UESTC/DOSE

[13] https://huggingface.co/datasets/issai/Turkish_Speech_Corpus/tree/main

[14] https://github.com/microsoft/MS-SNSD/tree/master