Türkçe TTS Sistemlerinin Geliştirilmesi için Dengeli Bir Veri Kümesi Hazırlama

Araştırma Makalesi/Research Article

Saadin Oyucu1*, Mustafa Sami Cücen2, Hüseyin Polat3

1Bilgisayar Mühendisliği Bölümü, Adıyaman Üniversitesi, Adıyaman, Türkiye
2Bilgisayar Mühendisliği Bölümü, Ostim Teknik Üniversitesi, Ankara, Türkiye
3Bilgisayar Mühendisliği Bölümü, Gazi Üniversitesi, Ankara, Türkiye

saadinoyucu@adiyaman.edu.tr, mustafasami.cucen@ostimteknik.edu.tr, polath@gazi.edu.tr
(Geliş/Received:08.08.2022; Kabul/Accepted:30.06.2023)
DOI: 10.17671/gazibtd.1159289

Özet— Konuşma sentezleme (TTS: Text-to-Speech) sistemleri insan-bilgisayar etkileşiminin önemli bir parçasıdır. TTS işleminde bir dizi metne karşılık gelen bir dizi spektrogram tahmin edilmektedir. Elde edilen spektrogram dizisi insanların duyabileceği ses dalga formuna dönüştürülmektedir. TTS sistemlerinin başarısı, geliştirme kaynaklarının yetersizliği nedeni ile farklı diller için aynı düzeyde değildir. Bir TTS sisteminin verimli şekilde geliştirilebilmesi için ulaşılabilir, büyük boyutlu bir konuşma veri kümesine ihtiyaç duyulmaktadır. Türkçe gibi kaynak yetersizliği olan diller için konuşma veri kümelerinin eksikliği, TTS sistemleri geliştirmenin önündeki en büyük engellerden biridir. Büyük boyutlu bir veri kümesi hazırlama oldukça zaman alan, zorlu ve maliyetli bir görevdir. Bu çalışmada, Türkçe TTS sistemlerinin geliştirilmesinde kullanılabilecek bir veri kümesi hazırlanmıştır. Daha önceden hazırlanan metin verisi, bir erkek konuşmacı tarafından İstanbul Türkçesi kullanılarak duygudan bağımsız olarak seslendirilmiştir. Metin verisi 109.826 kelime içermektedir. Seslendirilen konuşma verisi yaklaşık 12 saat 38 dakika 59 saniye uzunluğundadır ve 22.050 Hz. örnekleme frekansında kaydedilmiştir. Türkçe için hazırlanan bu veri kümesi daha önce İngilizce için hazırlanmış ve başarılı sonuçlar elde edilmiş "The LJ Speech Dataset" isimli veri kümesi ile karşılaştırılmış ve gelecekteki çalışmalar için öneriler sunulmuştur. Bu veri kümesi akademik düzeyde Türkçe TTS çalışmalarını teşvik etmek için hazırlanmıştır. Hazırlanan Türkçe veri kümesinin performans durumunu gözlemlemek için GlowTTS modeli bu veri kümesi kullanılarak eğitilmiştir. Eğitilen GlowTTS modeli ile bir Türkçe TTS sistemi geliştirilmiştir. Geliştirilen Türkçe TTS sistemi kullanılarak sentezlenen konuşmalar ile doğal konuşmaların karşılaştırılması sonucu 2,12'lik bir MOS-LQO değeri elde edilmiştir. Elde edilen ilk sonuçlar hazırlanan veri kümesinin Türkçe TTS sistemi geliştirme çalışmalarına etkin bir katkı sağladığını göstermektedir.

Anahtar Kelimeler— TTS, konuşma sentezleme, türkçe konuşma sentezleme, türkçe veri kümesi

Preparing A Balanced Corpus for Development of Turkish Speech Synthesis Systems

Abstract— Text-to-Speech (TTS) systems are an important part of human-computer interaction. In the TTS process, a series of spectrograms are predicted for a given text, which is then converted into waveforms that can be heard by humans. The success of TTS systems is not equal for different languages due to limited development resources. To efficiently develop a TTS system, a large, accessible corpus is needed. The lack of such corpuses, especially for languages with resource constraints such as Turkish, is one of the biggest obstacles to developing TTS systems. Creating a large corpus is a time-consuming, challenging, and costly task. In this study, a corpus was created that can be used in the development of Turkish TTS systems. The text data that was previously prepared was voiced by a male speaker using Istanbul Turkish, regardless of emotion. The text data contains 109,826 words. The voiced speech data is approximately 12 hours, 38 minutes, and 59 seconds long and was recorded at a sampling frequency of 22050 Hz This Turkish corpus was compared to "The LJ Speech Dataset," which was prepared for English and yielded successful results, and suggestions were made for future studies. This corpus was prepared to encourage academic-level Turkish TTS studies while avoiding academic plagiarism. In order to observe the performance of the prepared Turkish corpus, the GlowTTS model was trained using this dataset. A Turkish TTS system was developed with the trained GlowTTS model. A MOS-LQO value of 2.12 was obtained as a result of comparing the voice synthesized using the developed Turkish TTS system with the natural voice. Preliminary results show that the prepared corpus makes an effective contribution to the

Keywords—TTS, speech synthesis, turkish speech synthesis, turkish corpus

1. GİRİŞ (INTRODUCTION)

Konuşma sentezleme (TTS: Text-to-Speech) sistemleri temel olarak bir metnin sese dönüştürülmesini sağlayan teknolojiyi ifade etmektedir [1]. TTS sistemleri akustik, dilbilim, sinyal işleme ve istatistik gibi birçok farklı disiplinin birlikte kullanılması ile geliştirilmektedir. TTS sistemlerinden insan sesi doğallığına yakın sentetik konuşmaları üretmesi beklenmektedir. Başarı oranları giderek artan TTS sistemleri insan-makine etkileşimi, nesnelerin interneti, çağrı merkezleri, iletişim, sesli yanıt sistemleri ve eğitim alanında sesli kitap gibi farklı uygulamalarda kullanılmaktadır.

TTS alanındaki ilk uygulamalarda insan sesinin taklit edilmesi amacıyla farklı mekanik makine ve elektronik modeller geliştirilmiştir [2]. 1791'de Wolfgang Von Kempelen sadece harflerin değil tam sözcüklerin de üretilebilir olduğunu göstermiştir. Kempelen bir akustik konuşma makinesi geliştirmiştir [3]. Bilim insanları 1930'lu yıllara kadar Kempelen'in geliştirdiği makine üzerinde çalışmalar yapmıştır. 1930'lu yıllarda Bell Laboratuvarlarında konuşmayı otomatik olarak temel tonlarına ve titreşimlerine göre analiz edebilen bir ses kodlayıcı geliştirilmiştir [2]. Geliştirilen bu kodlayıcı sistem üzerinde TTS çalışmaları yapan Homer Dudley, Voder isimli ilk elektronik TTS makinesini geliştirmiştir [4]. Voder makinesi insan müdahalesi olmadan konuşma yapma yeteneğine sahip ilk sistem olarak bilinmektedir. Mekanik makinelerden elektronik sistemlere geçişten sonra Umeda ve arkadaşları tarafından 1968'de genel İngilizce metin okumayı gerçekleştiren ilk sistem geliştirilmiştir [5].

1970'lerin sonları 1980'lerin başlarında birçok TTS sistemi ticari amaçla üretilmiştir. Yazılımın yanında donanım çözümleri de içeren farklı TTS sistemleri bilgisayarlarda kullanılmaya başlanmıştır. DECtalk, Whistler, MBROLA gibi başarıları dikkate alınabilecek birçok TTS sistemi farklı diller için geliştirilmiştir. Türkçe TTS sistemi üzerine gerçekleştirilen çalışmalar 1990'lı yıllarda başlamıştır [6,7]. Türkçe TTS sistemleri ile ilgili olarak gerçekleştirilen çalışmalar veri kümesi hazırlama, dil modeli ve akustik model geliştirme üzerine yoğunlaşmıştır [6-8].

TTS sistemlerinin geliştirilmesi için eklemeli (birleştirmeli) sentezleme, formant (biçimlendirici) sentezleme, söyleyiş sentezleme, istatiksel parametrik sentezleme ve son yıllarda kullanılmaya başlayan Derin Öğrenme (DL: Deep Learning) tabanlı yöntemler mevcuttur.

Eklemeli TTS sistemleri temel olarak konuşma bilgisinden uygun birimlerin seçilmesini, seçilen birimleri ekleyen algoritmaları ve ekleme sınırlarını yumuşatmak için sinyal işleme çalışmalarını içermektedir [9]. Formant tabanlı TTS sistemleri, ses yolu aktarım fonksiyonunun, formant frekansları ve formant genlikleri benzetilerek üretilebilmesi üzerine gerçekleştirilen çalışmalardır [10]. Söyleyiş sentezlemeyi temel alan TTS sistemleri, insanların söyleyiş davranışının doğrudan modellenmesiyle konuşmayı sentezlemeye çalışmaktadır. Bu nedenle prensipte yüksek kaliteli konuşma sentezi üretmek için en başarılı yöntem olarak kabul edilmektedir. Ancak pratikte uygulanması en zor yöntemlerden biridir [11].

İstatistiksel parametrik TTS sistemlerinde konuşma sentezi için model tabanlı bir yaklaşım izlenmektedir. Eklemeli sistemlerin aksine model tabanlı yaklaşımda, birimleri depolamak yerine her bir birime karşılık gelen modeller bir havuzda saklanmaktadır. Model tabanlı yaklaşımda ölçeklendirilen konuşma örneklerinden elde edilen parametreler için gerekli modeller, istatistiksel yöntemlerle hazırlanmaktadır [12]. Türkçe gibi sondan eklemeli bir yapıya sahip diller için TTS geliştirme yöntemi olarak en uygun yöntemin eklemeli TTS geliştirme yöntemi olduğu belirtilmiştir [13]. Bu teknik için kullanılan veri kümeleri konuşma örneklerinden oluşmaktadır. Bu örnekler sözcükler, heceler, yarı heceler, ses birimleri, çift sesler veya üçlü seslerden oluşabilir. Örneklerin birim uzunluğu, sentezlenen konuşmanın doğallık ve anlaşılırlığını doğrudan etkilemektedir.

Doğallık ve anlaşılırlık üzerine gerçekleştirilen değerlendirmelerde eklemeli TTS sistemlerinin anlaşılırlık için daha iyi sonuçlar verdiği belirtilmiştir [6]. Ancak doğallık değerlendirmesinde ses birimleri arasındaki geçişlerde, konuşma doğallığında sorunlar yaşanmaktadır. Doğallık sorunun önüne geçmek için son yıllarda uygulanmaya başlanan DL tabanlı TTS sistemi geliştirme yöntemleri önerilmiştir. Doğal özelliklerin öğrenilmesinde etkili olduğu kanıtlanmış derin sinir ağları ile dil özelliklerinden elde edilen akustik özellikler doğrudan haritalanmaktadır [1]. Böylelikle dil özellikleri ile akustik özellikler arasında doğrudan bir bağ kurulmaktadır. DL temelli TTS sistemleri karmaşık modeller içermekte ve bu modellerin eğitilmesi sırasında dil özelliklerinin daha iyi öğrenilebilmesi için büyük veri kümelerine ihtiyaç duyulmaktadır.

DL tabanlı TTS yöntemlerinin geliştirilmesinde metin ve ses dosyalarının birebir eşleştirilmesi ile hazırlanan derlem tipindeki veri kümeleri kullanılmaktadır. Veri kümesinin büyüklüğü, konuşma sürelerinin uzunluğu, konuşmalarda geçen kelime ve benzersiz kelime sayısı geliştirilen sistemin başarısını doğrudan etkilemektedir.

İngilizce gibi zengin kaynaklara sahip dillerde TTS sistemi geliştirmek için veri kümesine erişmek kolaydır [14]. Ancak Türkçe TTS sistemi geliştirme sürecinde kullanılabilecek oldukça sınırlı küçük çaplı bazı veri kümeleri bulunmaktadır. Türkçe için yeterli büyüklükte ve iyi kaliteye sahip erişilebilir bir veri kümesi henüz mevcut değildir. TTS sistemi geliştirme sürecinde kullanılabilecek kaliteli bir veri kümesinin oluşturulması için bir konuşmacıdan çok fazla konuşma örneğinin alınması gerekir. Türkçe üzerine yapılan sınırlı veri kümesi çalışmalarında genellikle birden fazla konuşmacıdan birden fazla konuşma örneği alınmıştır [15-19]. Bu

farklılık konuşmaların aynı tonda ve özellikte modellenmesini zorlaştırmaktadır. Ayrıca Türkçe için daha önce geliştirilen veri kümelerinin çoğunluğu TTS görevi için değil daha çok konuşma tanıma görevi için uygundur. Yeterli kalitede ve büyüklükte Türkçe TTS veri kümesinin bulunmamasından dolayı bu çalışma kapsamında, Türkçe TTS sistemlerinin geliştirilmesinde kullanılabilecek doğallık ve anlaşılırlık başarımı yüksek, yeterli büyüklükte bir veri kümesi hazırlanmıştır. Hazırlanan veri kümesi içerisindeki her bir metin, kendisine karşılık gelen ses dosyası ile eşleştirilmiştir.

Bu çalışmanın birinci bölümü TTS sistemleri ile ilgili detaylı bilgiler verecek şekilde organize edilmiştir. İkinci bölümde, literatürde yer alan ve TTS sistemlerinde kullanılmak üzere geliştirilen farklı veri kümeleri detaylı olarak incelenmiştir. Üçüncü bölümde, hazırlanan Türkçe veri kümesi hakkında istatistiksel veriler analiz edilmiştir. Dördüncü bölümde deneysel sonuçlar verilmiştir. Son bölümde ise çalışma kapsamında elde edilen bulgular değerlendirilerek gelecekte yapılacak çalışmalar için öneriler sunulmuştur.

2. İLGİLİ ÇALIŞMALAR (RELATED STUDIES)

Literatürde farklı diller üzerine farklı yöntemler kullanılarak TTS sistemlerinin geliştirildiği görülmüştür. TTS çalışmalarının ilk dönemlerinde eklemeli sentezleme, formant sentezleme ve söyleyiş sentezleme yöntemleri kullanılmıştır. Daha sonraki dönemlerde istatiksel parametrik TTS sistemi geliştirme yöntemleri kullanılmıştır. Son yıllarda ise doğallık ve anlaşılırlık başarısını arttıran uçtan uca TTS sistemi geliştirme yöntemleri tercih edilmiştir. Uçtan uca TTS sistemi geliştirme yöntemleri ile yüksek kalite de dalga formları üretilerek başarı oranları arttırılmıştır. Ancak TTS sistemlerinin başarısı sadece uygulanan yönteme bağlı değildir. TTS sistemi geliştirilirken kullanılan veri kümesi, başarı oranları üzerinde büyük etki göstermektedir.

Formant sentezleme yöntemi birkaç parametrenin varyasyonu üzerinde çalıştığı için herhangi bir veri kümesine ihtiyaç duymamaktadır [20]. Söyleyiş sentezleme yöntemi, ses birimlerinin insan ses mekanizmasında nasıl oluşturulduğunun en hassas şekilde modellenmesini amaçlamaktadır. Ses üretim mekanizması içinde olan organların modellerini elde etmenin zorluğu nedeniyle bu organların davranışlarını yapay olarak modellemek zorlu bir görevdir [21]. Eklemeli sentezleme yöntemi ses birimlerini, belirli sinyal işleme teknikleriyle bir araya getirmeyi amaçlamaktadır. Bu nedenle birleştirilecek ses birimlerini içeren bir veri kümesine ihtiyaç duymaktadır. Bu yöntemde ses birimlerinin uzunluğu konuşmanın kalitesini doğrudan etkilemektedir. Daha uzun birimlerle, doğallık artarken daha az birleştirme noktasına ihtiyaç duyulmaktadır [21]. İstatiksel parametrik TTS yönteminde her bir birime karşılık gelen modeller bir havuzda saklanmaktadır. Model tabanlı yaklaşımda konuşma ölçeklendirilmekte ve konuşma ölçeklerinden elde edilen parametreler için gerekli modeller istatistiksel yöntemlerle hazırlanmaktadır [12]. Bu yöntem gerçekleştirilirken metin tabanlı bir veri kümesinin yanında akustik özellikleri de içeren bir veri kümesine ihtiyaç duyulmaktadır.

Derin öğrenme tabanlı TTS sisteminin başarısı model parametrelerinin yanı sıra sistemi eğitmek için kullanılan veri kümesine de bağlıdır. Bu nedenle veri kümesi TTS sistemleri için en iyi örnekleri içerecek şekilde hazırlanmalıdır. Veri kümesinde yanlış telaffuz edilmiş veya yanlış eşleştirilmiş ifadeler yer almamalıdır. Kayıttaki gürültülerin ve sessiz kısımların temizlenmiş olması gerekir. Elde edilen ses kayıtlarındaki farklılıklar bu yöntem ile gerçekleştirilen sistemlerin başarısını negatif yönde etkilemektedir. Konuşma parçası sürelerinin dağılımı belirli istatiksel yöntemler kullanılarak yapılmalıdır. Veri kümesi toplam uzunluğu konuşma sentezinin yapılacağı dile göre değişkenlik göstermektedir. Toplam uzunluğu kısa olan veri kümeleri aşırı öğrenmeyi önlemek ve modeli doğru bir şekilde eğitmek için yetersizdir. Hazırlanan veri kümesi ilgili dil için sık kullanılan sözcükler dışında, telaffuzu zor olan kelimeleri de içerecek şekilde hazırlanmalıdır [22].

Veri kümelerinde dikkat edilen özellikler konuşma dili, konuşma süreleri, konuşmacının cinsiyeti gibi bilgilerdir. [23,24]. Literatürde karşılaşılan veri kümeleri hakkında detaylı bilgiler Tablo 1'de verilmiştir. Derlem şeklinde hazırlanmış veri kümelerinde derlemin büyüklüğü, dilin kendi içindeki çeşitliliği (Örneğin; Türkiye'de konuşulan Türkçe ile Kuzey Kıbrıs Türkiye Cumhuriyeti'nde konuşulan Türkçe) gibi durumlar veri kümesinden elde edilebilecek bilgileri arttırmaktadır. İlgili çalışmalarda hazırlanan veri kümeleri konuşmanın uzunluğu, kelime veya cümle sayısı belirtilerek açıklanmıştır. Seslendirilen metin, alana özgü ise o metnin konusu belirtilmelidir. Li ve arkadaşlarının yaptığı çalışmada hedef TTS sistemi için tarımsal metinlerin seslendirilmesi gerçekleştirilmiştir [25]. Alana özgü bir çalışma yapılacak ise seslendirilmesi gereken metin ilgili alandan konuşmaları içermelidir. Böylelikle alana özgü ifadelerin istatistiksel olarak modellenmesi daha kolay olacaktır.

Genel konuşma alanına hitap eden TTS sistemleri için farklı dillerde birçok veri kümesi oluşturma çalışması gerçekleştirilmiştir. Arapça için "Arabic Speech Corpus" isimli bir veri kümesi Southampton Üniversitesi'nde Nawar Halabi tarafından hazırlanmıştır. Bu veri kümesi, profesyonel bir stüdyo kullanılarak Şam aksanı içeren Arapça konuşmaların seslendirilmesi ile kaydedilmiştir. Erkek bir konuşmacı tarafından yaklaşık 17.040 kelime seslendirilmiştir. Bu konuşmanın uzunluğu 3,7 saattir [26].

Tablo 1. Literatürdeki veri kümeleri için özellik/birim tablosu
(Feature/unit table for corpus in the literature)

Veri Kümeleri		**Özellikler (Birimler)**					
Referans	**Dil**	**Kelime Sayısı** *(Adet)*	**Benzersiz Kelime Sayısı** *(Adet)*	**Toplam Konuşma Kaydı Uzunluğu** *(Saat)*	**Konuşma Parçalarının Ortalama Uzunluğu** *(Saniye)*	**En Kısa Konuşma Parçası** *(Saniye)*	**En Uzun Konuşma Parçası** *(Saniye)*
[27]	İngilizce	10.045	2.974	1,08	3,43	1	7
[28]	İngilizce	225.715	13.821	23,92	6,57	1,11	10,10
[29]	İngilizce	782.815	14.564	73,35	8,5	1	25
[30]	Portekizce	71.358	13.311	10,47	10,37	0,67	50,08
[31]	Kazakça (Erkek)	320.200	42.400	57,1	8,3	0,8	55,9
	Kazakça (Kadın)	245.400	34.200	36,1	7,5	1	24,2
[32]	Baskça (Kadın)	30.901	8.583	7,44	6,77	2	15
	Baskça (Erkek)	26.383	8.030	6,6	7,25	3	14
[32]	Katalanca (Kadın)	24.385	6.568	5,4	8,38	3	19
	Katalanca (Erkek)	20.261	6.514	4,02	7,53	2	18
[32]	Galiçyaca (Kadın)	49.674	6.530	7,67	6,48	2	17
	Galiçyaca (Erkek)	15.462	4.336	2,64	7,19	3	20
[32]	Arjantin İspanyolcası (Kadın)	35.360	4.107	5,61	5,15	1	11
	Arjantin İspanyolcası (Erkek)	16.914	3.343	2,42	4,79	1	10
[32]	Şile İspanyolcası (Kadın)	16.591	3.279	2,84	5,9	2	11
	Şile İspanyolcası (Erkek)	25.168	4.171	4,31	5,89	1	12
[32]	Kolombiya İspanyolcası (Kadın)	22.228	4.460	3,74	5,68	1	13
	Kolombiya İspanyolcası (Erkek)	23.957	4.459	3,84	5,45	2	11
[32]	Peru İspanyolcası (Kadın)	23.806	4.278	4,35	6,18	2	14
	Peru İspanyolcası (Erkek)	27.547	4.268	4,87	6,01	2	14

Veri Kümeleri		Özellikler (Birimler)					
Referans	**Dil**	**Kelime Sayısı** *(Adet)*	**Benzersiz Kelime Sayısı** *(Adet)*	**Toplam Konuşma Kaydı Uzunluğu** *(Saat)*	**Konuşma Parçalarının Ortalama Uzunluğu** *(Saniye)*	**En Kısa Konuşma Parçası** *(Saniye)*	**En Uzun Konuşma Parçası** *(Saniye)*
[32]	Porto Riko İspanyolcası (Kadın)	6.092	1.738	1	4,23	2	12
[32]	Venezuela İspanyolcası (Kadın)	15.182	3.419	2,41	5,42	2	13
	Venezuela İspanyolcası (Erkek)	16.613	3.612	2,4	4,92	1	14
[32]	Guceratça (Kadın)	23.199	8.203	4,3	6,97	1	20
	Guceratça (Erkek)	21.518	7.818	3,59	6,3	1	17
[32]	Kannadaca (Kadın)	16.062	8.622	4,31	7,11	1	25
	Kannadaca (Erkek)	14.413	7.381	4,17	7,89	1	26
[32]	Malayalamca (Kadın)	12.581	5.713	3,02	5,17	1	15
	Malayalamca (Erkek)	12.749	6.407	2,49	4,43	1	12
[32]	Marathice (Kadın)	17.989	3.072	3,02	6,92	2	16
[32]	Tamilce (Kadın)	15.880	6.620	4,01	6,18	1	13
	Tamilce (Erkek)	13.545	66.159	3,07	5,66	1	17
[32]	Teluguca (Kadın)	11.286	4.218	2,73	4,28	1	14
	Teluguca (Erkek)	11.172	4.336	2,98	4,98	1	12
[32]	Yorubaca (Kadın)	15.880	4.113	2,06	3,9	1	11
	Yorubaca (Erkek)	14.242	3.835	1,97	4,2	1	12

Düşük kaynaklı Asya ve Afrika dilleri için Google tarafından çeşitli veri kümesi çalışmaları gerçekleştirilmiştir. Bu çalışmalarda 6 farklı dil için toplamda 8 veri kümesi geliştirilmiştir. Seylanca için 13 farklı kadın konuşmacı tarafından seslendirilen 3,38 saat uzunluğunda veri kümesi geliştirilmiştir. Bangladeş Bengalcesi için erkek konuşmacılar tarafından seslendirilen 2,94 saat uzunluğunda veri kümesi geliştirilmiştir. Hindistan Bengalcesi için erkek konuşmacılar tarafından seslendirilen 2,03 saat uzunluğunda veri kümesi geliştirilmiştir. Cava dili için 20 farklı erkek konuşmacı tarafından seslendirilen 3,47 saat uzunluğunda veri kümesi geliştirilmiştir. Cava dili için ek olarak 21 farklı kadın konuşmacı tarafından seslendirilen 3,52 saat uzunluğunda veri kümesi geliştirilmiştir. Kmerce için 17 farklı kadın konuşmacı tarafından seslendirilen 3,97 saat uzunluğunda veri kümesi geliştirilmiştir. Nepalce için 19 farklı kadın konuşmacı tarafından seslendirilen 2,8 saat uzunluğunda veri kümesi geliştirilmiştir. Sundaca için 22 farklı erkek konuşmacı tarafından seslendirilen 2,17 saat uzunluğunda veri kümesi geliştirilmiştir. Sundaca

için ek olarak 21 farklı kadın konuşmacı tarafından seslendirilen 3,21 saat uzunluğunda veri kümesi geliştirilmiştir [32]. Van Niekerk ve arkadaşlarının yaptıkları çalışmada Afrikanca, Sotho, Tsvana ve Xhosa dilleri için 4 farklı veri kümesi geliştirilmiştir. Afrikanca için 9 farklı kadın konuşmacı tarafından seslendirilen 3,32 saat uzunluğunda veri kümesi geliştirilmiştir. Sotho dili için 6 farklı kadın konuşmacı tarafından seslendirilen 3,22 saat uzunluğunda veri kümesi geliştirilmiştir. Tsvana dili için 8 farklı kadın konuşmacı tarafından seslendirilen 3,52 saat uzunluğunda veri kümesi geliştirilmiştir. Xhosa dili için 5 farklı kadın konuşmacı tarafından seslendirilen 3,1 saat uzunluğunda veri kümesi geliştirilmiştir [32,33].

Literatürde TTS sistemlerinin geliştirilmesinde kullanılabilecek İngilizce için hazırlanmış birçok veri kümesi bulunmaktadır. "CMU_ARCTIC" veri kümesi Carnegie Mellon Üniversitesi Dil Teknolojileri Enstitüsü tarafından oluşturulmuştur. Geliştirilen veri kümesinde Amerika, Kanada, İskoçya ve Hindistan aksanlarında hazırlanmış olan 4 farklı aksan için veri kümeleri mevcuttur. Amerikan İngilizcesi hem kadın hem de erkek konuşmacı tarafından seslendirilmiştir. Ancak diğer aksanlar için sadece erkek konuşmacı tarafından seslendirmeler gerçekleştirilmiştir. Bu veri kümelerinin her birinin 1.150 ifadeden oluştuğu belirtilmiştir. Bu veri kümelerinin toplam 10.000'in üzerinde kelimenin kullanıldığı, 3.000'e yakın benzersiz kelimenin kullanıldığı ve toplam kayıt uzunluğunun yaklaşık 1,08 saat olduğu görülmüştür [27,34].

İngilizce için hazırlanmış ve literatürdeki çalışmalarda da sıklıkla kullanılan veri kümelerinden bir diğeri de "LJSpeech Dataset"dir. LJ Speech Dataset, metin ve ses dosyalarının eşleştirilmesi şeklinde hazırlanmış derlem tipi veri kümesinin en temel örneklerindendir. LibroVox projesi kapsamında geliştirilen bu veri kümesi 7 farklı kitap bölümlerinin tek bir konuşmacının seslendirmesiyle hazırlanmıştır. Kadın bir konuşmacı tarafından seslendirilmiş olan bu veri kümesi, 13.821'i benzersiz olmak üzere toplam 225.715 kelimeden oluşmaktadır. Bu bilgilere ek olarak 1-10 saniye arasında uzunluklara sahip 13.100 konuşma parçasından oluşan veri kümesi toplamda 24 saat uzunluğundadır [28].

İngilizce üzerine hazırlanan bir diğer veri kümesi de "The World English Bible" isimli veri kümesidir. The World English Bible, Güney Amerika Birleşik Devletleri aksanlı bir erkek tarafından seslendirilmiş ve 73,35 saat uzunluğundadır [29]. Birleşik Krallık ve İrlanda'nın çeşitli yerel bölgelerinde konuşulan İngilizce aksanları için Google araştırmacıları tarafından veri kümesi çalışması yapılmıştır. Bu veri kümelerinin metin içeriği Vikipedi, Grant Fairbanks kamuya açık metni olan Rainbow Passage ve Alice Harikalar Diyarında isimli eserden hazırlanmıştır. Bunlara ek olarak yerel aksanlar için çeşitli yerel cümleler eklenmiştir. İrlanda aksanı için 3 erkek konuşmacı tarafından seslendirilen veri kümesinde 1.888'i benzersiz olan toplam 6.042 kelime mevcut olup 0,72 saat uzunluğundadır. Orta bölge aksanı için 2 kadın konuşmacı tarafından seslendirilen veri kümesi 1395'i benzersiz olan toplam 3.468 kelime içermektedir ve 0,47 saat uzunluğundadır. Orta bölge aksanı için ek olarak 3 erkek konuşmacının seslendirdiği 1.978'si benzersiz toplam 6.310 kelime içeren ve 0,73 saat uzunluğunda veri kümesi hazırlanmıştır. Kuzey bölge aksanı için 5 kadın konuşmacı tarafından seslendirilen veri kümesi 2.707'si benzersiz olan toplam 10.215 kelime içermektedir ve 1,37 saat uzunluğundadır. Kuzey bölge aksanı için ek olarak 14 erkek konuşmacının seslendirdiği 5.438'i benzersiz toplam 28.594 kelime içerip, 3,63 saat uzunluğunda veri kümesi hazırlanmıştır. İskoç aksanı için 6 kadın konuşmacı tarafından seslendirilen veri kümesi, 3.069'u benzersiz olan toplam 12.187 kelime içermektedir ve 1,58 saat uzunluğundadır. İskoç aksanı için ek olarak 11 erkek konuşmacının seslendirdiği 4.539'u benzersiz toplam 22.194 kelime içeren, 2,75 saat uzunluğunda veri kümesi hazırlanmıştır [32,35].

Literatürde farklı diller için farklı veri kümelerinin geliştirilmesi üzerine çalışmalar mevcuttur. Bu dillerden biri olan Estonca için hazırlanmış derlem tipi veri kümesi, kadın bir konuşmacı tarafından gazete metinlerinin seslendirilmesiyle elde edilmiştir. Toplam uzunluğu 1,08 saat uzunluğundadır [36,37]. Brezilya Portekizcesi için erkek bir konuşmacı tarafından seslendirilen 13.311'i benzersiz kelime olmak üzere toplam 71.358 kelime içeren ve 10,47 saat uzunluğunda olan bir veri kümesi geliştirilmiştir [30]. Almanca için Thorsten Müller tarafından 22,96 saat uzunluğunda bir veri kümesi hazırlanmıştır. Bu veri kümesi, profesyonel olmayan erkek bir konuşmacı tarafından seslendirilmiştir [38].

Japonca için "JSUT Corpus" isimli bir veri kümesi hazırlanmıştır. Japonca için hazırlanan veri kümesi profesyonel olmayan bir kadın konuşmacı tarafından seslendirilmiştir. Veri kümesi Vikipedi'den ve farklı derlemlerden elde edilen metinlerden faydalanılarak oluşturulmuştur. Veri kümesinin toplam konuşma uzunluğu 10,27 saattir [39,40]. Kazakça için "KazakhTTS" isimli veri kümesi hazırlanmıştır. KazakhTTS veri kümesinde seslendirilen metnin içeriği siyaset, iş, spor, eğlence gibi haber sitelerinden elde edilen çeşitli makalelerin filtrelenmesiyle oluşturulmuştur. KazakhTTS profesyonel bir kadın ve bir erkek olmak üzere iki konuşmacı seslendirmiştir. Kadın konuşmacı 44 yaşında ve 14 yıllık mesleki deneyime sahiptir. Kadın konuşmacı yaklaşık 34.200'ü benzersiz kelime olan 245.400 kelimeyi seslendirmiştir. Bu konuşmanın uzunluğu 36,1 saattir. Erkek konuşmacı ise 46 yaşında ve 12 yıllık meslek deneyimine sahiptir. Yaklaşık 42.400'ü benzersiz kelime olan 320.200 kelimeyi seslendirmiş ve bu konuşmasının uzunluğu 57,1 saattir [31,41]. Hintçe için kadın konuşmacı tarafından seslendirilen, 105.115 kelimeden oluşan 25,6 saat uzunluğunda olan bir veri kümesi geliştirilmiştir. Malayalamca için kadın konuşmacı tarafından

seslendirilen, 109.245 kelimeden oluşan 29,1 saat uzunluğunda bir veri kümesi geliştirilmiştir. Bengalce için erkek konuşmacı tarafından seslendirilen 104.891 kelimeden oluşan 22,3 saat uzunluğunda bir veri kümesi geliştirilmiştir [42].

Hindistan'ın çeşitli yerel bölgelerinde konuşulan diller için Google araştırmacıları tarafından veri kümesi çalışmaları yapılmıştır. Guceratça için 18 kadın konuşmacı tarafından seslendirilen, 8.203'ü benzersiz olan toplam 23.199 kelime içeren ve 3 saat uzunluğunda bir veri kümesi geliştirilmiştir. Guceratça için ek olarak 18 erkek konuşmacının seslendirdiği 7.818'i benzersiz olan toplam 21.518 kelime içeren ve 3,59 saat uzunluğunda bir veri kümesi daha geliştirilmiştir [32]. Kannadaca için 23 kadın konuşmacı tarafından seslendirilen veri kümesinde 8.622'si benzersiz olan toplam 16.062 kelime içeren ve 2,84 saat uzunluğunda bir veri kümesi geliştirilmiştir. Kannadaca için ek olarak 36 erkek konuşmacının seslendirdiği 7.381'i benzersiz olan toplam 14.413 kelime içeren ve 4,17 saat uzunluğunda veri kümesi geliştirilmiştir [32]. Malayalamca için 24 kadın konuşmacı tarafından seslendirilen, 5.713'ü benzersiz olan toplam 12.581 kelime içeren ve 3,02 saat uzunluğunda veri kümesi hazırlanmıştır. Malayalamca için ek olarak 18 erkek konuşmacının seslendirdiği 6.407'si benzersiz olan toplam 12.749 kelime içeren ve 2,49 saat uzunluğunda bir veri kümesi geliştirilmiştir [32]. Marathice için 9 kadın konuşmacı tarafından seslendirilen 3.072'si benzersiz olan toplam 17.989 kelime içeren ve 3,02 saat uzunluğunda veri kümesi geliştirilmiştir [32].

Dünya genelinde farklı diller için veri kümesi oluşturma çalışmaları gerçekleştirilmiştir. Ancak mevcut durumda Türkçe TTS sistemi geliştirebilmek için araştırmacıların kullanabileceği yeterli büyüklükte bir veri kümesi mevcut değildir. İstatiksel parametrik ve DL tabanlı TTS sistemi geliştirme yöntemlerinin kullanıldığı çalışmalarda ilgili çalışmaya özgü bazı Türkçe veri kümeleri hazırlanmıştır. Orhan ve Demiroğlu'nun yaptığı bir çalışmada iki kadın konuşmacı tarafından seslendirilmiş toplam 1.750 cümleden oluşan bir veri kümesi geliştirilmiştir [24]. Güner ve Demiroğlu'nun yaptığı çalışmada İstanbul Türkçesi konuşan kadın bir konuşmacının 500 kelimeyi seslendirdiği 70 dakika uzunluğunda bir veri kümesi hazırlanmıştır [43]. Bu veri kümelerinin İngilizce ve diğer diller için hazırlanmış çalışmalardaki veri kümesi uzunlukları dikkate alındığında çok yetersiz olduğu görülmektedir. Gökay ve Yalçın tarafından hazırlanmış bir başka çalışmada ise TTS sisteminin eğitimi için profesyonel bir konuşmacı tarafından seslendirilmiş 18 saatlik bir konuşma verisinin kullanıldığı belirtilmiştir [44]. Ancak belirtilen bu veri kümeleri araştırmacılar için ulaşılabilir değildir.

3. KONUŞMA SENTEZLEME İÇİN TÜRKÇE VERİ KÜMESİNİN HAZIRLANMASI (PREPARING OF TURKISH CORPUS FOR SPEECH SYNTHESIS)

Çok katmanlı yapay sinir ağlarını kullanan TTS sistemlerinde dil özelliklerinden akustik özelliklere doğrudan haritalama işlemi gerçekleştirmektedir. Konuşma sentezi için DL tabanlı yöntemleri benimseyen çok sayıda modelden biri de tamamen uçtan uca konuşma sentez modeli sunan ve kodlayıcı-kod çözücü mimarisini kullanarak giriş metninden mel spektrogramları üreten modellerdir. Mel spektrogramları üzerinde işlem yapmanın zor olması ve akış-tabanlı bir model kullanılması nedeniyle bu modeli eğitmek oldukça maliyetlidir [45]. Modelin başarısı eğitiminde kullanılacak veri kümesine doğrudan bağlıdır. Dolayısıyla TTS sistemi geliştirilmeden önce ilgili dil için uygun veri kümesinin oluşturulması gerekir.

Türkçe TTS sistemleri üzerine yapılan çalışmalar incelendiğinde hece tabanlı veri kümeleri ve difon tabanlı veri kümeleri hazırlamaya yönelik çalışmaların gerçekleştirildiği görülmüştür [46,47]. Ancak istatiksel parametrik TTS ve DL tabanlı TTS geliştirme yöntemleri için derlem şeklinde veri kümeleri hazırlanmıştır. Bu veri kümelerinde dikkat edilmesi gereken özellikler Tablo 2'de verilmiştir. Türkçe üzerine hazırlanmış derlemlerin konuşma tanıma sistemleri ve istatiksel parametrik TTS sistemleri üzerine uygulandığı görülmüştür [23, 44].

Tablo 2. Veri kümesi için özellik/birim tablosu
(Feature/unit table for corpus)

Özellik	**Birimi**
Konuşmanın Dili	Türkçe, İngilizce, İspanyolca vb.
Konuşmanın Süresi	Saat
Konuşmanın Kategorisi	Metin konusu
Konuşmacının Cinsiyeti	Kadın/Erkek
Konuşmacının Yaşı veya Yaş Düzeyi	Çocuk, yaşlı, 18 ve üzeri veya genç
Konuşmacının Şive Türü	İstanbul Türkçesi, Ege veya Karadeniz Şivesi
Konuşma Parçasının Ortalama Süresi	Saniye
Konuşma Parçasının Minimum Süresi	Saniye
Konuşma Parçasının Maksimum Süresi	Saniye
Benzersiz Kelime/Cümle Sayısı	Adet
Toplam Kelime/Cümle Sayısı	Adet

3.1. Genel Süreçler

Veri kümesi hazırlanırken konuşma sentezleme sisteminin hedeflerini ve kullanım amacını netleştirmek önemlidir. Amaca uygun içerik kaynaklarının seçilmesi gerekir. TTS sisteminin amacı ve kullanım alanına göre derlem tipinde hazırlanmış bir veri kümesinin içeriği günlük konuşmalar, edebi metinler, haber bültenleri

içeriği veya mantıksal bağlantısı olmayan bir dizi bağımsız cümle olabilir. Bu konuların çeşitlendirilerek özenle seçilmesi gerekir.

Veri kümesinin alanına uygun kaynaklar belirlendikten sonra veri kümesi tasarımına geçilmelidir. Bu aşamada, toplanan verilerin türü ve kapsamı belirlenmeye çalışılır. Tasarım sonucunda elde edilen toplam kelime sayısı ve benzersiz kelime sayısı gibi istatistiki bilgiler veri kümesinin daha dengeli bir yapıda olmasını sağlayacaktır. Özellikle üretken dil yapısına sahip Türkçe için daha fazla benzersiz kelimenin seslendirilmesi gerekir. Daha fazla farklı kelimenin seslendirilmesi fonemler arasındaki tahmin bağlantılarını güçlendirmektedir. Konuşmanın toplam uzunluğu, veri kümesi için önemli noktalardan biridir. Ayrıca her bir konuşma parçasının uzunluğu, geliştirilen model yapısında kullanılacak bacth size (bir yinelemede kullanılan eğitim örneklerinin sayısı) ve epoch (model eğitim adımlarının her biri) süresi üzerinde doğrudan etkiye sahiptir. Metin ve ses kaydı olarak birebir eşleştirilen konuşmalar kısa konuşma parçaları halinde saklanmalıdır.

Veri kümesinde yer alan konuşma içeriğinin yanı sıra konuşmacı hakkındaki bilgiler de geliştirilecek olan TTS sisteminin başarısını etkilemektedir. Konuşmacının cinsiyeti, ana dili ve kullandığı şive sentezlenen konuşmanın doğallığı ve anlaşılırlığına etki etmektedir. Literatürde aynı dil için hazırlanan veri kümelerinde farklı cinsiyete sahip konuşmacılardan faydalanılmıştır. Konuşmacının kullandığı şive türü sentezlenecek konuşmanın da aynı şive yapısına sahip olmasına neden olacaktır. Bu nedenle olabildiğinde şive kullanımından kaçınılmalıdır.

Kaydedilen veriler için temizlenme ve doğrulama işlemlerinin yapılması gerekmektedir. Bu aşamada, ses kayıtları gerçek kullanıcıların kontrolü ile değerlendirilmeli ve veri kümesindeki tutarsızlıklar veya hatalar düzeltilmelidir. Daha sonra veri kümesinde varsa sessizlik anlarının temizlenmesi gerekmektedir. Bu işlemin ana nedeni TTS sistemlerinin geliştirilmesinde kullanılan modellerde ses ve metin eşleştirmesinin daha net yapılmasını sağlamaktır. Ses-metin hizalamaları yapılırken ses kayıtlarının başında, sonunda ve içerisinde yer alan sessizlik anları model geliştirilirken gerçekleştirilen hizalamaları negatif yönde etkilemektedir.

3.2. Veri Kümesinin Hazırlanması

Genel süreçler göz önünde bulundurularak hazırlanan veri kümesi kullanım amacına uygun olarak hazırlanmıştır. Veri kümesi hazırlanırken Türkçe atasözleri, haber bültenleri, raporlar, kitap bölümleri, dini, edebi ve tarihi metinler, ülkelerin isimleri, Türkiye Cumhuriyeti'nin illeri, ilçeleri, çeşitli mahalle ve cadde isimleri, Türkiye Cumhuriyeti'nin çeşitli kurum ve kuruluşlarına ait isimler, sayılar ve mantıksal bağlantısı olmayan bir dizi bağımsız cümleler kullanılmıştır. Cümleler oluşturulurken ve seçilirken; Türkçe'nin, Ural-Altay dilleri grubuna bağlı sondan eklemeli yapısı göz önüne alınarak sona eklenen yapım ve çekim eklerine, ses değişimi (ses yumuşaması, ses sertleşmesi, ses düşmesi, ses eklemesi ve ulama gibi) kurallarına, sesteş kelimelere, düzenli veya devrik cümle yapılarına ve vurgulamalara dikkat edilmiştir.

Veri kümesi tasarımı kapsamında, konuşmacı seçimi ve toplanacak verilerde hedeflenen istatistiki veriler benzer veri kümeler göz önünde bulundurularak planlanmıştır. Bir konuşmacı seçiminde daha geniş konuşmacı sınıfına ulaşmak için en çok cinsiyet ve yaş dikkate alınmaktadır [48]. Farklı dillerde hazırlanan benzer veri kümelerinde [30,38] konuşmacı özelliklerine benzer şekilde İstanbul Türkçesi kullanan genç erkek konuşmacı tercih edilmiştir. Hazırlanan metin verisi daha önce profesyonel eğitim almış gönüllü bir konuşmacı tarafından seslendirilmiştir.

Konuşma kayıtları Audio Technica marka stüdyo tipi kardioid kondansatör özelliğine sahip bir mikrofon kullanılarak elde edilmiştir. Kayıt esnasında çevre gürültüsü ve ses patlamalarını azaltmak için bir pop filtresi kullanılmıştır. Bu mikrofonun tercih edilmesindeki en önemli özellikler kullanımının basit, gürültüsüz ve çok yönlü ses alabilmesidir. Kayıtların alınması için ofis ortamında ses yalıtımı yapılarak bir stüdyo hazırlanmıştır. Elde edilen veriler, tek kanallı, örnekleme hızı 22050 Hz ve bit derinliği 32-bit kayan olarak .wav formatında saklanmıştır.

3.3. Veri Ön İşleme ve Doğrulama Süreçleri

Literatürde Türkçe TTS sistemleri için hazırlanan veri kümelerinde, noktalama işaretleri ve kısaltmalardaki telaffuz farklılıkları, nümerik ifade kombinasyonlarındaki okuma alışkanlıkları ve ses değişimlerinin dikkate alınmadığı belirtilmiştir [49]. Bu nedenle çalışma kapsamında elde edilen konuşma kayıtları veri ön işleme ve doğrulama aşamasından önce 5 farklı kullanıcı tarafından kontrol edilmiştir. Elde edilen bütün kayıtlar 4 eşit parçaya ayrılarak kontrol amaçlı 4 farklı kullanıcıya sunulmuştur. Kullanıcılar bütün kontrolleri gerçekleştirdikten sonra son kontrolcü, kontrol edilen bütün kayıtları birleştirmiş ve tekrar kontrol işlemini gerçekleştirmiştir.

Veri ön işleme ve doğrulama aşaması için bütün ses kayıtları tek tek kontrol edilerek kaliteleri değerlendirilmiştir. Veri kümesindeki tutarsızlıklar veya hatalar tespit edilip düzenlenmiş veya veri kümesinden çıkarılmıştır. Bu kapsamda 100'e yakın ses dosyası veri kümesinden çıkarılmıştır. Veri kümesinin ön işleme ve doğrulama aşamasında konuşma içerisindeki sessizlik anları konuşma kaydından silinmiştir. Ses kayıtlarındaki sessizlik durumları kontrol edilerek ilgili kayıt içerisinden çıkarılmıştır.

Konuşma kayıtları içerisindeki sessizlik durumlarının tespiti ve kaldırılması için öncelikle durağan ve aralıklı konuşmalardaki sessizlik bilgisi tespit edilmiştir. Bu işlem için her bir konuşma parçasındaki ses dalgalarının örneklem frekansı ve genliğinden yararlanılmıştır. Belirli bir frekans ve genlik değerinin aşağısında yer alan konuşma bilgisinde sessizliğin olduğu varsayılmıştır. Eşik değerinin altına düşüldüğü ilk andan itibaren arka planda bir zaman sayacı tutulmuştur. Bu sayaç sessizlik ile karşılaşıldığı ilk andan itibaren sürekli olarak arttırılmıştır. Sayaç belirli bir sınır değerine ulaştığında sessizlik bilgisinin var olduğu tespit edilmiştir. Son işlem olarak sayacın başlama zamanından bitiş zamanına kadar olan içerik konuşma kaydından silinmiştir.

Otomatik olarak gerçekleştirilen sessizliklerin tespiti ve kaldırılması işleminin yanı sıra gerçek kullanıcılar tarafından sessizlik kontrolü gerçekleştirilmiştir. Kontrol işlemlerinde ses kayıtlarının başlangıcı ve sonlarındaki sessizlik anları silinmiştir. Böylelikle geliştirilecek olan modellerde sessizlik anları dikkate alınmayacaktır. Metin-ses hizalamalarında zaman değişimi olmayacak ve model parametreleri herhangi bir sapmaya uğramayacaktır.

Kontrol işlemlerinde sadece sessizlik anları değil aynı zamanda metinlerin okunuşları da kontrol edilmiştir. Gerçek kullanıcılar tarafından gerçekleştirilen bu işlem sayesinde metinlerin telaffuzlarının doğruluğu kontrol edilmiştir. Sadece kelimelerin okunuşlarına değil aynı zamanda rakamlar, sayılar ve kısaltmaların da okunuşlarına da dikkat edilmiştir. Yanlış veya eksik telaffuzlar veri kümesinden çıkarılmıştır.

Türkçedeki kısaltmalar üzerine hem eğitim hem de eğitim sonrasında giriş metni üzerinde çalışabilecek ek bir liste hazırlanmıştır. Bu listede yer alacak kısaltmalar için Türk Dil Kurumu'nun belirlediği kısaltmalardan yararlanılmıştır. Kısaltmaların veri kümesine eklenmesinde yazılışları değil okunuşları dikkate alınmıştır. Böylelikle ses-metin hizalamasında karakter bazlı değil fonem temelli hizalamada kullanılabilecek bir veri sunulmuştur.

Nümerik ifade kombinasyonları için veri kümesinde seslendirilen metnin içine çeşitli nümerik ifadeler eklenmiştir. Nümerik ifadelerin veri kümesinde eklenmesi işleminde nümerik ifadelerin yazılışı değil okunuşları dikkate alınmıştır. Ayrıca para birimlerini ve bu birimlerin okunuşları veri kümesine eklenmiştir. Böylelikle hesaplama ve para birimleri ile ilgili konuşma sentezlerinin daha gürbüz sonuçlar vereceği ön görülmektedir.

Veri kümesinin doğrulama işlemlerinde son işlem olarak noktalama işaretleri gözden geçirilmiştir. Metinlerin okunuşlarındaki vurguları net olarak ortaya koyabilmek ve konuşma sentezinde gerekli vurguları elde edebilmek için noktalama işaretleri tek tek kontrol edilmiştir. Bu kontrol işlemlerinde özellikle virgül ve nokta işaretlerine dikkat edilmiştir. Tüm bu süreçlerden sonra elde edilen veri kümesine ait istatiksel özellikler Tablo 3'te verilmiştir.

3.4. Veri Kümesine Ait İstatistiksel Özellikler

Veri boyutunu ve çeşitliliğini belirlemek, kelime dağarcığını zenginleştirmek ve farklı metinlerin doğru bir şekilde sentezlenmesini sağlamak amacıyla önemli bilgiler sunan istatistiksel özellikler Tablo 3'te verilmiştir.

Tablo 3. Türkçe veri kümesi için özellik birim tablosu
(Feature unit table for Turkish corpus)

İstatiksel Özellikler	**Türkçe Veri Kümesi**
Toplam Kelime Sayısı	109.826 Adet
Benzersiz Kelime Sayısı	35.050 Adet
Toplam Karakter Sayısı	745.011 Adet
Toplam Konuşma Uzunluğu	12 Saat 38 Dakika 59 Saniye
Toplam Konuşma Parçası Sayısı	8.480 Adet
Ortalama Konuşma Parçası Uzunluğu	5,19 Saniye
Minimum Konuşma Parçası Uzunluğu	0,54 Saniye
Maksimum Konuşma Parçası Uzunluğu	9,85 Saniye
Konuşma Parçası Başına Düşen Ortalama Kelime Sayısı	12,95 Adet
Konuşmacının Cinsiyeti	Erkek
Konuşmacının Yaş Düzeyi	Genç
Konuşmacının Ağız Türü	İstanbul Türkçesi

Tablo 3'teki verilen konuşma parçalarının uzunluklarına ait çeşitli istatistikler TTS sisteminin ortalama bir konuşma süresine göre optimize edilip edilmediğini belirlemek, farklı uzunluktaki konuşmalara nasıl tepki vereceğini anlamak ve sınırları belirlemek için detaylı bilgiler içermektedir.

Tablo 4. Kelime frekansı analizi sonuçları
(Result of word frequency analysis)

Kelime	**Tekrarlama Sayısı**
bir	2.144
kimlik	574
olarak	557
bin	473
yüz	424
on	357
olan	337
tarafından	292
ilgili	289
dokuz	287

Hazırlanan veri kümesi içerisinde yer alan metinler hakkında genel bir fikir edinebilmek için kelime frekansı analizi gerçekleştirilmiştir. Gerçekleştirilen analiz sonucunda elde edilen en sık kullanılan 10 kelime Tablo 4'te verilmiştir.

Tablo 4'te verilen kelime frekansı analizi sonuçlarına göre en sık kullanılan kelimenin "bir" olduğu görülmüştür. Kelime frekansı analizi, sık kullanılan kelimelerin sentezlenmesinde doğru vurgu ve doğal akıcılık sağlamak için önemlidir. Sık kullanılan kelimelerin doğru bir şekilde sentezlenmesi, TTS sisteminin genel anlaşılabilirliğini ve performansını artıracağı düşünülmektedir. Aynı zamanda, kelime frekansı analizi, TTS sisteminin dil modellerinin ve sentezleme stratejilerinin geliştirilmesinde kullanılabilir.

Kelime frekansına ek olarak kelime başına düşen karakter sayısı hesaplanmıştır. Gerçekleştirilen işlemler sonucunda kelime başına düşen ortalama karakter sayısı 5,84 olarak belirlenmiştir. Elde edilen bulgular, hazırlanan veri kümesi ile geliştirilecek olan bir Türkçe TTS sisteminin kelime dağarcığı hakkında detaylı bilgiler vermektedir.

4. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

Yapılan araştırmalarda Türkçe dili için eklemeli TTS sistemlerinin anlaşılırlık için daha iyi sonuçlar verdiği görülmüştür. Ancak aynı yöntemin kullanıldığı TTS sistemlerinde ses birimleri arasındaki geçişlerde ve konuşma doğallığında sorunlar yaşanmaktadır. Doğallık sorunun önüne geçmek için DL tabanlı uçtan uca TTS yöntemleri tercih edilmektedir. DL temelli yaklaşımların başarılı olabilmesi için büyük boyutta metin ve ses dosyalarının eşleştirilmesi şeklinde hazırlanan veri kümesine ihtiyaç duyulmaktadır.

Metin ve ses dosyalarının eşleştirilmesi şeklinde hazırlanmış veri kümelerinin en iyi örneklerinden biri olan "The LJ Speech Dataset" ile yapılan İngilizce TTS geliştirme çalışmalarında MOS değeri 4,52 olarak elde edilmiştir. İngilizcede doğal insan konuşması için MOS değerinin 4,58 olduğu göz önüne alındığında başarılı sonuçların elde edildiği gözlemlenmiştir [45]. İlgili çalışmalar incelendiğinde hazırlanan veri kümesi ile geliştirilecek olan Türkçe TTS sistemlerinin başarılı sonuçlar vereceği tahmin edilmektedir. Tablo 5'te İngilizce TTS çalışmalarında başarılı sonuçlar veren The LJ Speech Dataset'e ait özellik birim tablosu bu çalışmada hazırlanan veri kümesi ile karşılaştırılarak verilmiştir. Ek olarak konuşma özellikli uygulamaları eğitmek için kullanabileceği açık kaynaklı, çok dilli bir konuşma veri kümesi olan Mozilla Common Voice Corpus 13.0 ile karşılaştırmalar yapılmıştır.

Tablo 5'te verilen özellik birim tablosu incelendiğinde İngilizce için başarılı sonuçlar veren The LJ Speech Dataset veri kümesinin bu çalışmada hazırlanan veri kümesi ile benzer özelliklere sahip olduğu görülmüştür. Toplam konuşma süresi İngilizce için hazırlanan veri kümesinde daha uzundur. Toplam konuşma süresi, veri kümesinin boyutu ve içeriği hakkında fikir edinebilmek için önemlidir. Toplam konuşma süresi, TTS sistemi için gerekli olan tahmin modelinin daha fazla spektrogram örneği görmesini sağlar. Bu durum TTS modelinin kelime dağarcığını arttırabileceği gibi aynı zamanda konuşma örneklerindeki temel frekans farklılıklarının daha iyi modellenmesini sağlar.

Tablo 5'te Türkçe olarak erişime sunulan Common Voice Corpus 13.0 veri kümesi hakkında detaylı bilgilerde verilmiştir. Bu veri kümesi birden fazla kullanıcıdan birden fazla örnek alınarak oluşturulmuştur. Akustik model geliştirmek ve akustik bilgi edinmek için kullanılması mümkündür. Ancak TTS sistemlerinde kullanılabilecek tek kullanıcıdan çok fazla örneği içermemektedir. Mevcut durumda Common Voice Corpus 13.0 veri kümesinde 1.775 kayıtlı konuşmacı mevcuttur [50].

Tablo 5. The LJ Speech Dataset veri kümesi için özellik birim tablosu
(Feature unit table for The LJ Speech Dataset)

İstatiksel Özellikler	**The LJ Speech Dataset (İngilizce)**	**Common Voice Corpus 13.0 (Türkçe)**	**Türkçe Veri Kümesi**
Toplam Konuşma Parçası Sayısı	13.100 Adet	97.249 Adet	8.480 Adet
Toplam Kelime Sayısı	225.715 Adet	453.403 Adet	109.826 Adet
Toplam Karakter Sayısı	1.308.678 Adet	2.787.432 Adet	745.011 Adet
Toplam Konuşma Uzunluğu	23 Saat 17 Dakika 55 Saniye	98 Saat 8 Dakika 57 Saniye	12 Saat 38 Dakika 59 Saniye
Ortalama Konuşma Parçası Uzunluğu	6,57 Saniye	3,63 Saniye	5,19 Saniye
Minimum Konuşma Parçası Uzunluğu	1,11 Saniye	0.46 Saniye	0,54 Saniye
Maksimum Konuşma Parçası Uzunluğu	10,10 Saniye	10,82 Saniye	9,85 Saniye

Konuşma Parçası Başına Düşen Ortalama Kelime Sayısı	17,23 Adet	4,66 Adet	12,95 Adet
Benzersiz Kelime Sayısı	13.821 Adet	61.830 Adet	35.050 Adet

Hazırlanan Türkçe TTS veri kümesinin uygulanabilirliğini test etmek amacıyla bir TTS sistemi geliştirilmiştir. Geliştirilen Türkçe TTS sisteminde GlowTTS [51] kullanılmıştır. Bu çalışma sonucunda elde edilen sentezlenmiş ses ile doğal ses arasındaki ses kalitesi nesnel olarak ölçülmüştür. Nesnel bir ölçüm işlemi için ViSQOL (Virtual Speech Quality Objective Listener) [52] kullanılmıştır. ViSQOL, iki konuşma sinyali arasındaki benzerlik özelliklerini çıkarmakta ve bu özellikleri kullanarak sinyaller arasındaki farkı hesaplamaya çalışmaktadır. Gerçekleştirilen karşılaştırmalar sonucunda Ortalama Görüş Puanı-Dinleme Kalitesi Hedefi (MOS-LQO: Mean Opinion Score- Listening Quality Objective) [53] değeri elde edilmektedir.

Hazırlanan veri kümesi her biri 261 adım olan 200 çevrimlik model ile GlowTTS kullanılarak eğitilmiştir. Bu eğitim sonucunda yapılan deneylerde MOS-LQO değeri 5 üzerinden 2,12 olarak ölçülmüştür. Gerçekleştirilen TTS geliştirme işlemi dile özgü ek bir çalışma yapılmadan uçtan uca TTS geliştirme yapıları kullanılarak bir Türkçe TTS sisteminin geliştirilebilir olduğunu göstermiştir.

5. SONUÇ VE ÖNERİLER (CONCLUSIONS AND RECOMMENDATIONS)

Bu çalışmada hazırlanan veri kümesi kullanılarak gelecekte daha kapsamlı Türkçe TTS sistemlerinin geliştirilmesi planlanmaktadır. Türkçe TTS sisteminin geliştirilmesi aşamasında kullanılabilecek konuşmanın doğallığı ve anlaşılabilirliği üzerinde önemli bir etkiye sahip DL tabanlı GlowTTS gibi birçok araç ve yöntem mevcuttur.

Sonraki çalışmalarda Tacotron + WaveNet ve Tacotron 2 + HiFi-GAN tabanlı yöntemler kullanılarak bir Türkçe TTS sisteminin geliştirilmesi planlanmaktadır. Tacotron 2, WaveNet ve HiFi-GAN çalışmalarına dayanarak, mevcut TTS sistemlerinin insan seviyesinde kaliteye ulaşıp ulaşamayacağı LJSpeech veri kümesinde test edilmiş ve başarılı sonuçlar alınmıştır [45].

FastSpeech 2 + HiFi-GAN, Glow-TTS + HiFi-GAN, GradTTS + HiFi-GAN sistemleri üzerine yapılan çalışmalar [54], İngilizce konuşmaların yüksek kalitede sentezlenebileceğini göstermiştir. Bu çalışma kapsamında elde edilen veri kümesi, Türkçe için uygulanabilirliğini yapılan küçük eğitim sonucu alınan ortalama sonuç ile göstermiştir. İleride İngilizce gibi başarılı TTS çalışmalarında kullanılan yöntemler temel alınarak Türkçe için bir TTS sistemi geliştirilecektir. Gerçekleştirilecek çalışmalar ile veri kümesinin farklı yöntemler üzerindeki başarısı net olarak ortaya çıkarılacaktır. Ayrıca veri kümesinin kullanımının yaygınlaştırılması sağlanarak TTS alanında gerçekleştirilen Türkçe çalışmaların arttırılması hedeflenmiştir.

TEŞEKKÜR (THANKS)

Bu çalışma Türkiye Bilimsel ve Teknolojik Araştırma Kurumu tarafından desteklenmiştir (Proje no: 121E479)

KAYNAKLAR (REFERENCES)

[1] Y. Ning, S. He, Z. Wu, C. Xing, L. J. Zhang, "Review of deep learning based speech synthesis", Appl. Sci., 9(19), 1–16, 2019.

[2] S. Lemmetty**, Review of speech synthesis technology**, Yüksek Lisans Tezi, Helsinki University of Technology, Department of Electrical and Communications Engineering, 1999.

[3] H. Dudley, T. H. Tarnóczy, "The Speaking Machine of Wolfgang von Kempelen", J. Acoust. Soc. Am., 22 (2), 151–166, 1949.

[4] H. Dudley, "The Carrier Nature of Speech", Bell Syst. Tech. J., 19 (4), 495–515, 1940.

[5] N. Umeda, R. Teranishi, "The Parsing Program for Automatic Text-to-Speech Synthesis Developed at the Electrotechnical Laboratory in 1968", IEEE Trans. Acoust., 23 (2), 183–188, 1975.

[6] A. E. Yilmaz, "Türkçe Metinden Konuşma Sentezleme Uygulamaları İçin Bir Veri Sözlük Seti ve Yazılım Çerçevesi", Gazi Üniversitesi Mühendislik Mimar. Fakültesi Derg., 24 (4), 735–744, 2009.

[7] İ. Y. Özüm, **A Speech Synthesis System for Turkish Language Based on the Concetanation of Phonemes Taken from Speaker**, Yüksek Lisans Tezi, Middle East Technical University, Graduate School of Natural and Applied Sciences, 1993.

[8] B. Eker, **Turkish Text To Speech System**, Yüksek Lisans Tezi, Bilkent University, The Department of Computer Engineering, 2002.

[9] R. A. Khan, J. S. Chitode, "Concatenative Speech Synthesis: A Review", Int. J. Comput. Appl., 136 (3), 1–6, 2016.

[10] Y. Tabet, M. Boughazi, "Speech synthesis techniques. A survey", **7th Int. Work. Syst. Signal Process. their Appl. WoSSPA 2011**, 67–70, 2011.

[11] M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il, N. Mastorakis, "An overview of text-to-speech synthesis techniques", **Int. Conf. Commun. Inf. Technol. - Proc**., 84–89, 2010.

[12] D. Govind, S. R. M. Prasanna, "Expressive speech synthesis: A review", Int. J. Speech Technol., 16 (2), 237–260, 2013.

[13] R. Aşlıyan, K. Günel, "Türkçe metinler için hece tabanlı konuşma sentezleme sistemi", **Akademik Bilişim 2008**, Çanakkale, Türkiye, 31–38, 2008.

[14] H. Zen, V. Dang, R. Clark, Y. Zhang, Y. Jia, Z. Chen, Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech", **Conference of the International Speech Communication Association,** Graz, Avusturya, 15-19 Eylül, 2019.

[15] O. Salor, B. Pellom, T. Ciloglu, K. Hacioglu, M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language", **International Conference Spoken Language Processing,** Denver, Colorado, Amerika Birleşik Devletleri, 16-20 Eylül, 2002.

[16] O. Salor, T. Ciloglu, K. Hacioglu, M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language", **International Conference Spoken Language Processing,** Denver, Colorado, Amerika Birleşik Devletleri, 16-20 Eylül, 2002.

[17] O. Salor, B. Pellom, T. Ciloglu, M. Demirekler, "Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition", Computer Speech Language, 21(4), 580-593, 2007.

[18] E. Arisoy, D. Can, S. Parlak, H. Sak, M. Saraclar, "Turkish broadcast news speech and transcripts", IEEE Transactions on Audio Speech and Language Processing, 17(5), 874 – 883, 2009.

[19] İnternet: Türkçe Ulusal Derlemi (TUD) – Turkish National Corpus (TNC), https://www.tnc.org.tr/tr/, 05.04.2023.

[20] M. Jalil, F. A. Butt, A. Malik, "A survey of different speech synthesis techniques", **2013 Int. Conf. Technol. Adv. Electr. Electron. Comput. Eng. TAEECE 2013**, 204–207, 2013.

[21] İ. B. Uslu, "Metinden Konuşma Sentezleme", **TMMOB Elektrik Mühendisleri Odası Ankara Şubesi Haber Bülteni**, 11–14, 2010.

[22] A. Dunaev, **A Text-to-Speech System Based on Deep Neural Networks**, Lisans Tezi, KIT Department of Informatics, Institute for Anthropomatics and Robotics (IAR), Interactive Systems Labs (ISL) Karlsruhe Institute of Technology, 2019.

[23] B. S. Gürler, **Türkçe Konuşma Tanıma Sistemleri İçin Bir Konuşma Veritabanı**, Yüksek Lisans Tezi, Gazi Üniversitesi, Elektronik-Bilgisayar Eğitimi Anabilim Dalı, 2014.

[24] M. C. Orhan, C. Demiroğlu, "Konuşmacı Aradeğerlemeli SMM Tabanlı Metinden Konuşma Sentezleme Sistemi", **2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011**), 781–784, 2011.

[25] X. Li, D. Ma, B. Yin, "Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system", Comput. Electron. Agric., 180, 1–10, 2021.

[26] N. Halabi, **Modern Standard Arabic Phonetics for Speech Synthesis,** Doktora Tezi, University of Southampton, Faculty of Physical Sciences and Engineering School of Electronics and Computer Science, 2016.

[27] İnternet: Festvox, CMU_ARCTIC Databases, http://festvox.org/cmu_arctic/, 23.04.2022.

[28] İnternet: The LJ Speech Dataset, https://keithito.com/LJ-Speech-Dataset/, 23.04.2022.

[29] İnternet: Kaggle, The World English Bible, https://www.kaggle.com/datasets/bryanpark/the-world-english-bible-speech-dataset?select=transcript.txt, 23.04.2022.

[30] E. Casanova vd., "TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese", Lang. Resour. Eval., 2022.

[31] İnternet: Papers With Code, KazakhTTS Dataset, https://paperswithcode.com/dataset/kazakhtts, 23.04.2022.

[32] İnternet: openslr.org, https://www.openslr.org/, 23.04.2022.

[33] D. Van Niekerk vd., "Rapid development of TTS corpora for four South African languages", **Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH**, 2178–2182, 2017.

[34] J. Kominek, A. W. Black, "The CMU Arctic Databases for Speech Synthesis", **Proc. ISCA Work. Speech Synth**., 223–224, 2004.

[35] I. Demirşahin, O. Kjartansson, A. Gutkin, C. Rivera, "Open-source Multi-speaker Corpora of the English Accents in the British Isles", **Proc. 12th Language Resources and Evaluation Conference (LREC 2020),** 6532- 6541, 2020.

[36] İnternet: Meta-Share, Estonian Emotional Speech Corpus, https://metashare.ut.ee/repository/browse/estonian-emotional-speech-corpus/4d42d7a8463411e2a6e4005056b40024a19021a316b54b7fb707757d43d1a889/, 23.04.2022.

[37] R. Altrov, H. Pajupuu, "Estonian Emotional Speech Corpus: theoretical base and implementation", **4th International Workshop on Corpora for Research on Emootion Sentiment & Social Signals ES3 2012**, 50–53, 2012.

[38] İnternet: T. Müller and D. Kreutz, Thorsten-Voice- 'Thorsten-21.02-neutral' Dataset, https://zenodo.org/record/5525342, 23.04.2022

[39] İnternet: Papers With Code, JSUT Corpus Dataset, https://paperswithcode.com/dataset/jsut-corpus, 23.04.2022.

[40] R. Sonobe, S. Takamichi, H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis", **ICASSP2018**, 2017.

[41] S. Mussakhojayeva, A. Janaliyeva, A. Mirzakhmetov, Y. Khassanov, H. A. Varol, "KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset", **Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH**, 3511–3515, 2021.

[42] N. Srivastava, R. Mukhopadhyay, K. R. Prajwal, C. V Jawahar, "IndicSpeech: Text-to-Speech Corpus for Indian Languages", **Proc. 12th Language Resources and Evaluation Conference (LREC 2020),** 6417- 6422,2020.

[43] E. Guner, C. Demiroglu, “A small footprint hybrid statistical/unit selection text-to-speech synthesis system for agglutinative languages”, **ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc**., 4537–4540, 2012.

[44] R. Gokay ve H. Yalcin, “Improving Low Resource Turkish Speech Recognition with Data Augmentation and TTS”, **16th Int. Multi-Conference Syst. Signals Devices, SSD 2019**, 357–360, 2019.

[45] J. Shen vd.., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions", **2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 4779-4783, 2018.

[46] H. Tora, İ. B. Uslu, T. Karamehmet, “Implementation of Turkish text-to-speech synthesis on a voice synthesizer card with prosodic features”, Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering, 18(3), 584-594, 2017.

[47] I. B. Uslu, H. K. İlk, “A rule based perceptual intonation model for Turkish text-to-speech synthesis”, **2012 20th Signal Processing and Communications Applications Conference (SIU)**, Muğla, Türkiye, 18-20 Nisan, 2012.

[48] T. Schultz, **Speaker Classification I** , C. Müller, 4343, Springer, Berlin, Heidelberg, 2007.

[49] İ. Sel, D. Hanbay, M. Karabatak, “Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi”, Elektr.ve Bilgi. Sempozyumu 2011, 273–276, 2011.

[50] İnternet: Common Voice Mozilla, https://commonvoice.mozilla.org/tr/datasets, 05.04.2023.

[51] J. Kim vd., “Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search”, ArXiv, abs/2005.11129, 2020.

[52] A. Hines vd., “ViSQOL: an objective speech quality model”, Journal on Audio, Speech, and Music Processing, 2015(13), 2015.

[53] C. Sloan vd., “A. Objective assessment of perceptual audio quality using ViSQOLAudio”, IEEE Trans. Broadcast, 63, 693-705, 2017.

[54] X. Tan vd., “NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality”, ArXiv, abs/2205.04421, 2022.