T.C.

OSTİM TEKNİK ÜNİVERSİTESİ

ÜRETKEN YAPAY ZEKA DERSİ

YZM 308 PROJE RAPORU

Hazırlayan

Ayşegül KARTAL 220212014

ANKARA 2025

Öğr. Adı ve Soyadı	Ayşegül KARTAL	**Öğrenci No**	220212014
PROJE BAŞLIĞI	**Kendi Sesimle Fransızca Konuşan Video Oluşturma: TTS ve Görüntü Eşleme Uygulaması**		

Proje Özeti:

Bu proje kapsamında, kişisel ses kayıtlarımı kullanarak kendi sesimle Fransızca konuşuyormuş gibi görünen bir video oluşturmak amaçlanmıştır. Bunun için metinden-konuşmaya (TTS), dil çevirisi ve dudak senkronizasyonu gibi üretken yapay zekâ yöntemleri entegre edilmiştir. Tacotron 2 modeli ile kişiselleştirilmiş ses sentezi hedeflenmiş, ardından Fransızcaya çevrilen metinler ile Wav2Lip kullanılarak senkronize bir video elde edilmeye çalışılmıştır. Veri toplama ve ön işleme adımları başarıyla tamamlanmış; ancak model eğitimi ve çıktı oluşturma aşamaları teknik kısıtlamalar nedeniyle tamamlanamamıştır. Bu çalışma, çok dilli kişisel medya üretimi alanında önemli bir potansiyele sahiptir.

Giriş:

Günümüzde yapay zekâ destekli ses sentezi ve görüntü eşleme teknolojileri, çok dilli içerik üretiminde önemli bir yer edinmiştir. Bu proje kapsamında amaç, Türkçe olarak seslendirdiğim 120 cümlelik bir veri setini kullanarak, kendi sesimle Fransızca konuşuyormuş gibi görünen bir video oluşturmaktır. Proje, metinden-konuşmaya (TTS), ses dönüştürme, çeviri ve dudak senkronizasyonu gibi aşamaları içermektedir. Bu sayede, kişinin kendi ses kimliği korunarak farklı bir dilde içerik üretmesi hedeflenmektedir. Bu sayede, özellikle dil bariyerinin olduğu alanlarda daha etkili, erişilebilir ve doğal iletişim olanakları doğmaktadır.

Literatür Taraması:

Bu çalışmada kullanılan yöntemler daha önce birçok araştırmada başarıyla uygulanmış ve farklı amaçlarla değerlendirilmiştir. Literatürde öne çıkan bazı çalışmalar şunlardır:

- **Tacotron 2** (Shen et al., 2018): Karakter tabanlı metinleri doğal ve akıcı bir biçimde konuşmaya dönüştürmeyi başaran bir TTS modelidir. Özellikle kişisel ses klonlama için transfer öğrenme ile özelleştirme yapılabilmektedir.
- **Wav2Lip** (Prajwal et al., 2020): Girdi sesi ile herhangi bir yüz videosunu senkronize ederek dudak hareketlerini gerçeğe çok yakın şekilde oluşturabilmektedir. Haber spikerleri, dil öğrenme videoları ve film dublajı gibi alanlarda başarıyla kullanılmıştır.

- **Voice Cloning** üzerine yapılan çalışmalar (Jia et al., 2018), sadece birkaç dakikalık veriyle kişiye özel ses üretiminin mümkün olduğunu göstermektedir.
- **Multilingual TTS ve AI dubbing** sistemleri, özellikle Netflix, YouTube ve eğitim platformlarında içerikleri çok dilli sunmak amacıyla kullanılmaktadır.

Bu teknolojiler bir araya getirilerek kullanıcıya özel, çok dilli ve gerçekçi görsel-işitsel içeriklerin oluşturulması mümkün hâle gelmiştir.

Projenin Önemi:

Bu projenin önemi birkaç açıdan ele alınabilir:

- **Kişiselleştirilmiş İçerik Üretimi:** Bir kişinin başka bir dilde kendi sesiyle konuşmasını sağlamak, içerik üretiminde gerçekçilik ve bağlılık sağlar.
- **Erişilebilirlik:** İşitme veya görme engelli bireyler için dudak okuyarak veya sesle desteklenen içeriklerin sunulması önemlidir.
- **Eğitim ve Dil Öğrenimi:** Ana dili dışında içeriklere doğal sesle ulaşmak, eğitim süreçlerinde etkili olabilir.
- **Çok Dilli Yayıncılık:** Aynı video içeriğini farklı dillere uyarlayarak daha geniş kitlelere ulaşmak mümkündür.

YÖNTEMLER:

Bu projenin gerçekleştirilmesi için aşağıdaki adımlar planlanmıştır:

1. Veri Toplama ve Hazırlık

- 120 Türkçe ses kaydı kendi sesimle kaydedildi.
- Her ses kaydına ait metinler `metadata.csv` adlı dosyada saklandı.
- Kayıtlar `Tacotron2_Sounds` adlı bir Google Drive klasöründe organize edildi.

2. TTS Modeli Eğitimi (Tacotron 2 ile)

- Tacotron 2 modeli, kişisel ses kayıtlarımla özelleştirilerek eğitilmek istendi.
- Eğitim süreci Google Colab ortamında yürütülmeye çalışıldı.
- Ses verilerinin uygun formata dönüştürülmesi, spectrogram oluşturulması ve model parametrelerinin ayarlanması gibi işlemler planlandı.

3. Metin Çevirisi

- Türkçe cümleler, Fransızcaya çevrilerek hedef metinler elde edildi.
- Bu adımda Google Translate veya DeepL gibi otomatik çeviri sistemleri kullanıldı.

4 .Dudak Senkronizasyonu (Wav2Lip)

- Wav2Lip modeli ile bir yüz videosu üzerine hedef Fransızca metnin seslendirmesi senkronize edilerek dudak hareketleri oluşturulacaktı.
- Bu sayede, kişinin Fransızca konuşuyormuş gibi görünen gerçekçi bir video hedeflendi.

BULGULAR

Projede veri toplama ve ön hazırlık süreçleri başarıyla tamamlandı. Ancak, ses sentezi ve dudak senkronizasyonu aşamalarında çeşitli teknik engellerle karşılaşıldı:

- **Model Eğitimi Sorunları:** Tacotron 2 modelinin eğitimi için gerekli olan GPU kaynakları yetersiz kaldı veya eğitim süreci hatalarla kesildi.
- **Uyumsuzluklar:** Wav2Lip ile dudak senkronizasyonu sağlanması için gereken ses dosyalarının formatı ve modelin istekleri tam olarak karşılanamadı.
- **Zaman Kısıtı:** Eğitim süreci ve model çıktılarının oluşturulması zaman alıcı olduğundan, proje teslim süresine yetiştirilemedi.

Bu sebeplerle proje planlanan tamamlanmış çıktı olan "kendi sesimle Fransızca konuşan video" oluşturma aşamasına geçilemedi.

SONUÇ – TARTIŞMA

Bu proje, çok dilli yapay zeka uygulamalarında ses ve görüntü sentezlemenin birleştiği yenilikçi bir fikre dayanmaktadır. Veri toplama ve ön işleme süreçleri başarılı şekilde gerçekleştirilmiş olsa da, teknik zorluklar ve zaman kısıtı sebebiyle nihai çıktıya ulaşılamamıştır. Gelecekte şu geliştirmelerin yapılacağı amaçlanmaktadır:

- Model eğitimi için daha güçlü GPU kaynaklarının kullanılması,
- Veri ön işleme ve model uyumluluğunun daha dikkatli kontrol edilmesi,
- Daha fazla zaman planlaması yapılarak her adımın tamamlanması.

Bu çalışma, eksik kalmış olsa da ileriye dönük projeler için sağlam bir temel oluşturmuştur. Ses sentezi, kişiselleştirilmiş TTS ve görüntü senkronizasyonu gibi alanlarda yapılacak geliştirmelerle daha başarılı sonuçlara ulaşmak mümkündür.

KAYNAKÇA:

- Shen, J., Pang, R., Weiss, R. J., et al. (2018). *Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions*. Google AI.

- Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). *Wav2Lip: Accurately Lip-syncing Videos In The Wild*. ACM Multimedia.

- Jia, Y., Zhang, Y., Weiss, R. J., et al. (2018). *Transfer learning from speaker verification to multispeaker text-to-speech synthesis*. NeurIPS.

- Google Translate API.

- DeepL Translator.