UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

Aluno: Mário Sérgio Maduro Santana
Orientador: João Batista Florindo

Modelo de inteligência artificial para transição entre imagens: uma aplicação no tratamento da afasia

Campinas
Junho de 2024

Mário Sérgio Maduro Santana

Modelo de inteligência artificial para transição entre imagens: uma aplicação no tratamento da afasia

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do professor João Batista Florindo.

1 Introdução

No dia 26 de maio de 2022, Marcelo de Ornelas Santana, o pai do autor deste projeto, teve um acidente vascular cerebral (AVC) enquanto trabalhava, ficando com sequelas motoras e com Afasia - disfunção na capacidade de comunicação em geral, que é gerada por uma lesão no hemisfério esquerdo do cérebro. No início, a dificuldade na linguagem era completa, sendo que a escrita, a fala, a compreensão e até mesmo os gestos foram perdidos.

Após muito treinamento, os gestos começaram a fazer sentido, mas a escrita e a fala continuaram a ser um grande problema. Apesar disso, com a resiliência de Marcelo em seus estudos diários e depois de muito tratamento fonoaudiológico, ele começou a conseguir escrever algumas palavras-chave para expressar algo que ele queria. Além de tudo isso, a capacidade de compreensão melhorou consideravelmente, apesar de existirem muitas dificuldades.

Outro ponto a se considerar é que o Marcelo consegue repetir as palavras que falamos através da audição e da leitura labial. Por esse motivo surgiu a pergunta: Será que é possivel criar uma espécie de tradutor "texto-Movimento Labial"?

A resposta para esta pergunta é positiva, sendo que neste relatório iremos utilizar a rede FILM - *Frame Interpolation for Large Motion* - para superar este desafio.

2 O que é Afasia?

De acordo com o Hospital Israelita Albert Einstein, **afasia** é "uma disfunção que diminui a capacidade de uso da linguagem, da fala e prejudica a comunicação", a qual é gerada por lesões no hemisfério esquerdo do cérebro. Como consequência dessa lesão, "a pessoa não consegue se expressar verbalmente da mesma forma que faria antes, bem como apresenta dificuldades na compreensão da linguagem verbal, escrita e mesmo na capacidade de escrever". As principais causas dessas lesões cerebrais ocorrem devido a um "Acidente Vascular Cerebral (AVC), tumores cerebrais, doenças degenerativas (a Doença de Alzheimer, por exemplo) ou impactos na cabeça que acometem o hemisfério esquerdo do cérebro, ou as regiões frontais e temporais à esquerda".

Além do que foi dito anteriormente, minha experiência com meu pai me mostrou que os afásicos podem apresentar jargões na fala, os quais consistem em palavras (existentes ou não) que o afásico fala repetidamente com diferentes entonações, como se estivesse falando normalmente. Outrossim, apesar da grande dificuldade na escrita, após muito treinamento e tratamentos fonoaudiológicos, meu pai começou a conseguir escrever palavras-chave como uma forma de comunicação com as pessoas.

Por conta desse fato, a criação de um aplicativo o qual recebe uma palavra e a traduz para movimentos labiais é muito interessante, visto que esse artifício ajuda o afásico a planejar os movimentos da boca durante a fala de alguma palavra.

Neste projeto, faremos a 1ª etapa da confecção do aplicativo: iremos criar um meio de gerar movimentos labiais realistas. As próximas etapas para fechar o aplicativo serão desenvolvidas em um projeto de extensão futuro.

3 Apresentando o Problema

Dadas 2 imagens de "boquinhas", as quais representam, separadamente, um fonema do Português do Brasil, queremos gerar uma terceira imagem, que representa o *frame* de transição entre as 2 primeiras imagens.

Perceba que, por meio da solução para o problema supracitado, conseguiremos gerar uma animação de uma boca real falando qualquer sílaba do português do Brasil. Além disso, concatenando diversas animações, podemos gerar uma animação realista para, praticamente, qualquer palavra do nosso idioma.

À fim de solucionarmos o que foi proposto, utilizamos a rede neural FILM - *Frame Interpolation for Large Motion*, a qual, dadas 2 imagens quase duplicadas, gera um 3º *frame* intermediário que interpola as 2 imagens iniciais.

4 Rede FILM: Frame Interpolation for Large Motion

4.1 Informações Gerais

A rede FILM foi criada a partir de uma parceria entre pesquisadores do Google Research e da Universidade de Washington. Mais informações, tais como o artigo da rede e os códigos do GitHub, podem ser encontradas em: https://film-net.github.io/.

Cabe ressaltar que, neste projeto, utilizamos os pesos pré-treinados pelos pesquisadores que criaram a rede. Por esse motivo, não foi necessário um treinamento personalizado com os nossos próprios dados de treinamento.

Outrossim, uma observação importante para toda esta Seção (4) é que todas as explicações sobre a Rede FILM foram, fortemente, baseadas em seu artigo original.

4.2 Arquitetura da Rede

Como se observa na Figura (1), dadas 2 imagens de entrada $\boldsymbol{(I_0, I_1)}$, a rede FILM sintetiza uma imagem intermediária $\boldsymbol{\hat{I}_t}$, com tempo $t \in (0, 1)$. Isto é, se $\mathcal{M}(I_0, I_1)$ é a função que representa a rede neural, então:

$$\hat{I}_t = \mathcal{M}(I_0, I_1).$$

OBS.: Nosso interesse, neste projeto, é obter o rótulo $\hat{I}_{0,5}$, dado que queremos gerar um frame intermediário justamente no tempo médio entre as duas imagens de entrada I_0 e I_1.

Cabe dizer que a rede FILM possui 3 estágios principais, são eles:

- Extração de *Features* Compartilhadas;
- Estimativa de Fluxo de Movimento;
- Estágio de fusão;

Nas próximas subseções, serão explicados cada um desses estágios.

4.2.1 Extração de Features Compartilhadas

Como é possivel ver na Figura (1), as primeiras camadas da rede FILM são compostas por uma pirâmide de características, a qual é comumente utilizada para a extração de *features* de imagens. Note que cada uma das imagens, I_0 e I_1, são fornecidas como entradas para a rede em pirâmides de características distintas.

A fim de padronizarmos a notação para falar sobre uma determinada posição da pirâmide de características, vamos definir a profundidade $d \in [1, 4]$ e o nível $L \in [1, 7]$ da pirâmide como na Figura (2).

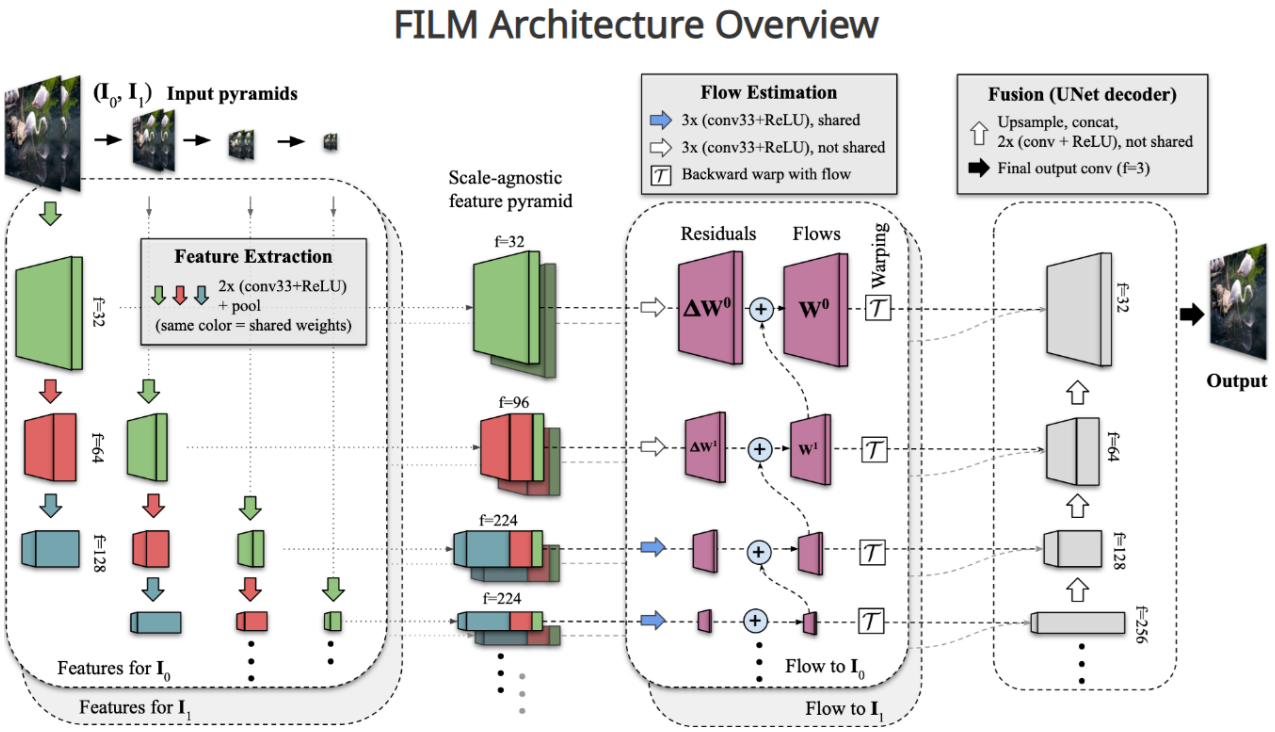

Figura 1: Arquitetura da rede FILM
Fonte:https://film-net.github.io/

Note, na Figura (2), que, em cada nível L, as mesmas profundidades d estão representadas pela mesma cor. Ainda é importante convencionar que o mapa de características presente na posição (d, L) das pirâmides referentes às imagens de entrada I_0 e I_1 é denotado, respectivamente, por $f_0^{L,d}$ e $f_1^{L,d}$, em que $f_i^{L,d} = \mathcal{H}^d(I_i^L)$ (para $i = 0$ ou 1), com $\mathcal{H}^d$ representando uma pilha de convoluções para cada profundidade d.

Apesar de as pirâmides de características serem boas extratoras de *features*, elas possuem, segundo o artigo da rede FILM, duas dificuldades:

1. Pequenos objetos com movimento rápido desaparecem em níveis mais grosseiros da pirâmide (Obs.: Os níveis grosseiros são as colunas da pirâmide de características que recebem imagens com resoluções baixas);

2. O número de pixels é drasticamente menor em níveis grosseiros da pirâmide, o que significa que há menos pixels para a extração de características em imagens com grandes movimentos.

Para resolver os dois problemas anteriores, os pesquisadores decidiram compartilhar os pesos das convoluções realizadas em uma mesma profundidade d de diferentes níveis L considerados. Dessa forma, partindo da intuição, considerada no artigo da rede, de que "grandes movimentos em escalas mais detalhadas são equivalentes a pequenos movimentos em escalas mais grosseiras", o compartilhamento de pesos permitiu o aumento da quantidade de pixels disponíveis para a supervisão de grandes movimentos. Além disso, como as convoluções possuem pesos compartilhados para diferentes escalas das imagens de entrada, as *features* capturadas em uma certa escala podem ser aproveitadas em outras escalas, o que melhora a obtenção das *features* de escalas mais grosseiras.

Após a extração de características, como podemos ver nas Figuras (1) e (3), os mapas de *features* $f_i^{L,d}$ obtidos nas pirâmides são concatenados da seguinte forma: $F_i^L = (f_i^{L-3,d=4}, f_i^{L-2,d=3}, f_i^{L-1,d=2}, f_i^{L,d=1})$, em que $i = 0$ ou 1.

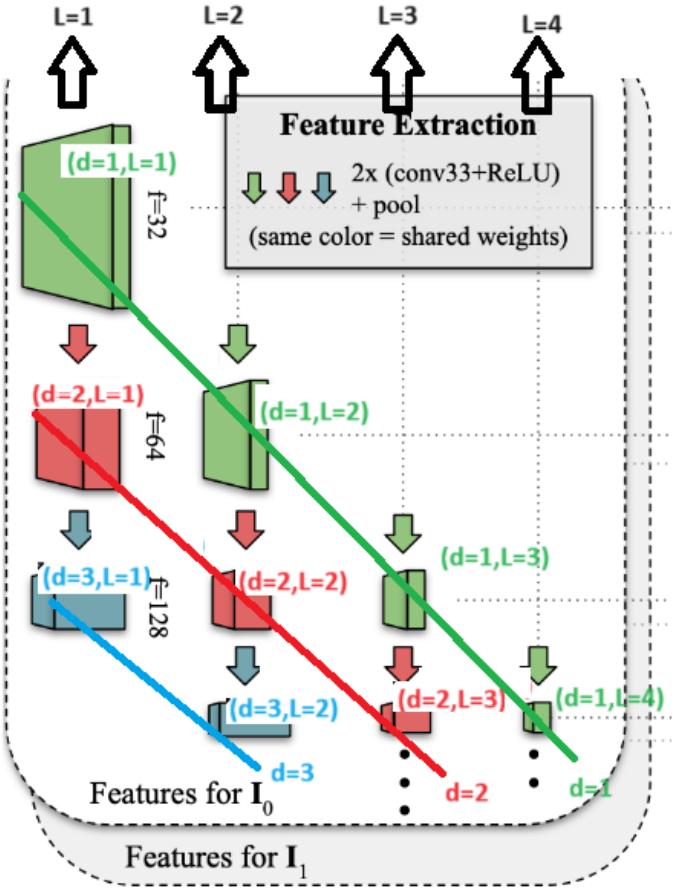

Figura 2: Profundidades d e níveis L da pirâmide de características. Fonte:https://film-net.github.io/

Após essa concatenação, as *features* resultantes servem como entrada para a próximo estágio: estimativa de movimento.

4.2.2 Estimativa de Fluxo de Movimento

Como o próprio nome diz, a função do estágio atual é estimar o movimento, entre os dois *frames* de entrada I_0 e I_1, a partir dos *features* extraídos durante o estágio de extração de características. Cabe acrescentar que neste ponto começa a ficar difícil de encontrar uma intuição convincente que justifique as operações realizadas.

A estimativa de movimento é iniciada no nível $L = 7$ utilizando as seguintes operações matemáticas:

$$\hat{F}^L_{t\leftarrow 0} = \tau(F^L_0, (W^{L+1}_{t\rightarrow 0})_{\times 2}) \tag{1}$$

$$W^L_{t\rightarrow 0} = (W^{L+1}_{t\rightarrow 0})_{\times 2} + \mathcal{G}^L(F^L_1, \hat{F}^L_{t\leftarrow 0}) \tag{2}$$

$$W^8_{t\rightarrow 0} = 0 \tag{3}$$

$$\hat{F}^L_{t\leftarrow 1} = \tau(F^L_1, (W^{L+1}_{t\rightarrow 1})_{\times 2}) \tag{4}$$

$$W^L_{t\rightarrow 1} = (W^{L+1}_{t\rightarrow 1})_{\times 2} + \mathcal{G}^L(F^L_0, \hat{F}^L_{t\leftarrow 1}) \tag{5}$$

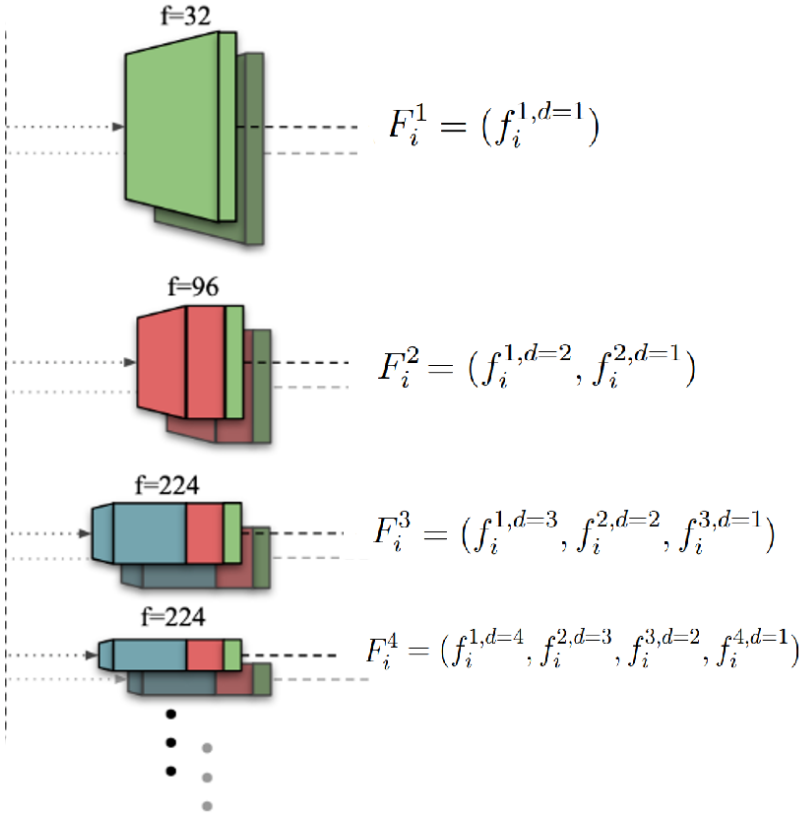

Figura 3: Concatenação das *features* obtidas. Fonte:https://film-net.github.io/

$$W_{t\to 1}^{8} = 0 \tag{6}$$

Notação:

- $W_{t\to 0}^{L}$: Representa o fluxo de movimento que leva o quadro intermediário do instante t para o quadro em $t = 0$ no nível L.
- $\hat{F}_{t\leftarrow 0}^{L}$: Representa o mapa de características do tempo $t = 0$ deformado de tal forma que ele se alinhe com o mapa de características do instante intermediário t no nível L.
- $W_{t\to 0}^{L}$: Representa o fluxo de movimento que leva o quadro intermediário do instante t para o quadro em $t = 1$ no nível L.
- $\hat{F}_{t\leftarrow 1}^{L}$: Representa o mapa de características do tempo $t = 1$ deformado de tal forma que ele se alinhe com o mapa de características do instante intermediário t no nível L.
- $(\cdot)_{\times 2}$: Representa um *up-sampling* bilinear, que é utilizado para aumentar a resolução.
- $\mathcal{G}^{L}$: É uma pilha de convoluções que serve para estimar o resíduo.
- τ Representa uma operação de reamostragem bilinear.

Após as estimativas anteriores, podemos obter a pirâmide de características no tempo intermediário t por meio da deformação reversa das pirâmides de carcteristicas nos tempos $t = 0$ e $t = 1$, isto é:

$$F_{t\leftarrow 0}^{L} = \tau((F_0^L, I_0^L), W_{t\to 0}^{L}) \tag{7}$$

$$F_{t\leftarrow 1}^{L} = \tau((F_1^L, I_1^L), W_{t\to 1}^{L}) \tag{8}$$

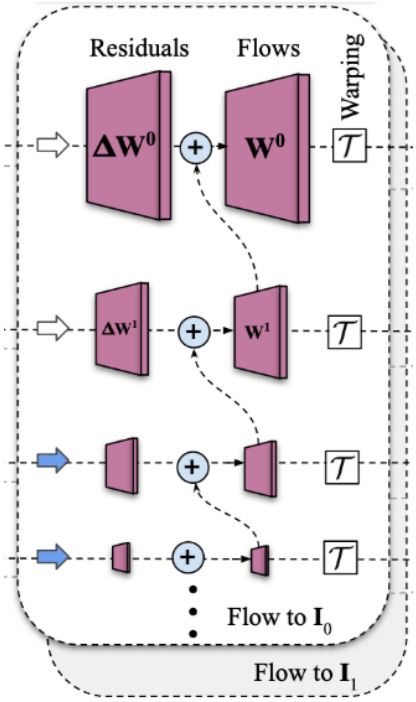

Figura 4: Estágio de Estimativa de Movimento. Fonte:https://film-net.github.io/

4.2.3 Estágio de Fusão

No estágio atual, a saída do último nível do estágio anterior (estimativa de fluxo de movimento) passa por convoluções e é concatenada à saída do penúltimo nível. O resultado dessas operações é, novamente, passado por convoluções e concatenados com os *features* do antepenúltimo nível. Essas operações são repetidas até o nível $L = 1$ do estágio de fusão. Por fim, através de uma nova convolução obtemos a imagem de saída $\hat{I}_t$. Veja a Figura 5 para uma melhor compreensão do que foi dito.

5 Metodologia e Resultados

5.1 Fotos das "Boquinhas"

Antes da fase de captura de fotos, o autor deste projeto procurou por imagens de alfabeto fonético (fotos das "boquinhas") na internet. Embora esse conteúdo esteja presente na web, inclusive para venda, todos os pedidos de autorização para o uso das imagens em um aplicativo - de código aberto e totalmente gratuito - foram negados.

Por esse motivo, as fotos utilizadas neste trabalho são da Elizabeth Aparecida Freitas Maduro, mãe do autor deste projeto.

5.2 Rodando a rede FILM

Como foi dito na seção de apresentação do problema (Seção 3), fizemos todos os arranjos 2 a 2 de todos os fonemas possíveis e rodamos a rede FILM várias vezes, sempre salvando o vídeo de resultado a cada iteração. Veja na Figura (6) um trecho de código que deixa claro o que foi dito.

Após rodar o código da Figura (6), temos uma pasta com os vídeos das animações das transições de cada fonema possível.

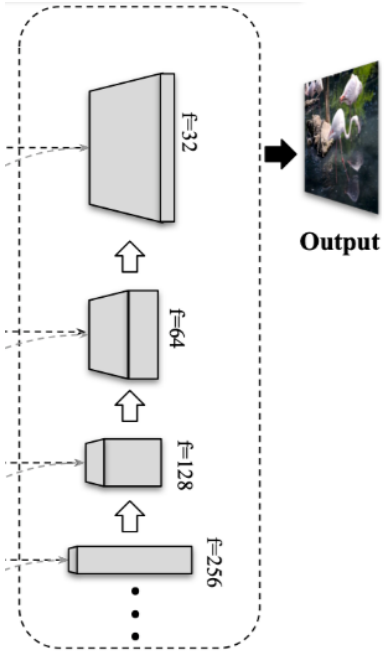

Figura 5: Estágio de Fusão. Fonte:https://film-net.github.io/

```
#Rodando a rede para todos os pares Possiveis
for img1 in file_list:
  for img2 in file_list:
    if img1!=img2:
      path1=f"/content/gdrive/MyDrive/Bocas_Fillm/{img1}"
      path2=f"/content/gdrive/MyDrive/Bocas_Fillm/{img2}"
      # Carregando as imagens
      image1 = load_image1(path1,224,266)
      image2 = load_image1(path2,224,266)
      # Entrada da rede
      input = {
      'time': np.expand_dims(time, axis=0),
      'x0': np.expand_dims(image1, axis=0),
      'x1': np.expand_dims(image2, axis=0)
      }
      # Rodando a rede Film
      mid_frame = model(input)
      #Salvando um vídeo da animação
      frames = [image1, mid_frame['image'][0].numpy(), image2]
      media.write_video(f'/content/gdrive/MyDrive/Video_Bocas_2FPS/{img1[:-4]}{img2[:-4]}.mp4', frames, fps=2, qp=18)
```

Figura 6: Rodando a rede FILM para todos os pares de fonemas possíveis

5.3 Obtendo os Resultados

No momento temos uma pasta com vídeos de todas as animações possíveis. Logo, para conseguirmos os movimentos labiais de uma palavra específica do português, basta concatenar as animações. Por exemplo, se quisermos uma animação da palavra "CAVALO", podemos concaternar as animações "CA", "AV","VA","AL" e "LO". Cabe ressaltar que, ao concatenar os vídeos, devemos acelerar os "vídeos de transição", isto é, no caso da palavra "CAVALO", devemos acelerar os vídeos das animações "AV" e "AL". Veja um vídeo no youtube dos movimentos labiais da palavra "CAVALO", "GATO" e "GELO", que foram concatenados manualmente: https://www.youtube.com/watch?v=PtMxyvt8yvc. **OBS.:** Após alguns testes, decidimos rodar os "vídeos de transição" com velocidade 3x.

Como se observa no vídeo acima, ainda é possivel realizar melhorias nos vídeos gerados, visto que falta um bom caminho para eles ficarem perfeitos. Além disso, como os "vídeos de transição" são rodados em uma velocidade distinta dos vídeos restantes, para automatizarmos a concatenação dos vídeos, é necessário realizar a separação silábica das palavras, dado que dessa forma é possível identificar quais serão as transições.

6 Próximos Passos

O próximo passo do projeto atual é aprimorar os vídeos gerados fazendo um possível *fine tuning* da rede FILM com os nossos próprios dados. Como também criar um meio de fazer a separação silábica das palavras da língua portuguesa, pois a qualidade dos movimentos labiais depende da velocidade em que as transições são rodadas.

7 Conclusão

O projeto atual cumpriu seu objetivo inicial, que propunha encontrar uma forma de gerar um tradutor Texto-para-Movimento Labial. Porém, como foi dito na seção anterior, ainda há muito trabalho a ser feito.

8 Referências

[1] Paper: "FILM: Frame Interpolation for Large Motion"; Autores:Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru e Brian Curless.

[2] Site: https://www.tensorflow.org/hub/tutorials/tf_hub_film_example; Título: Frame interpolation using the FILM model.