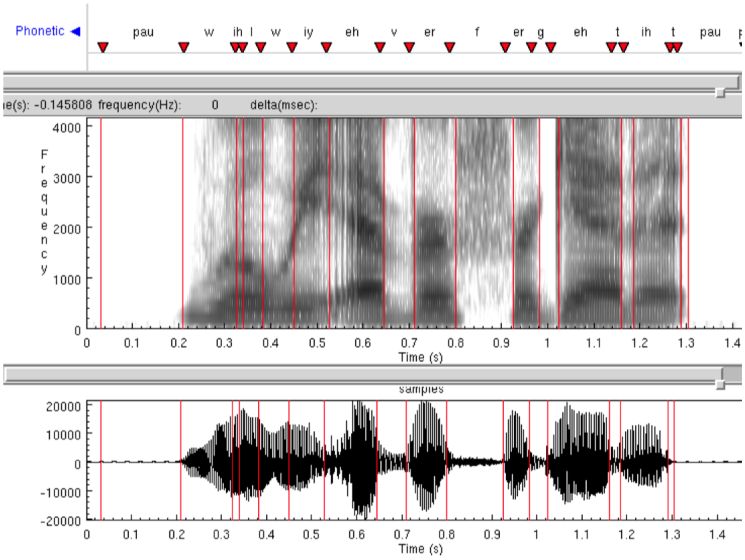

Figure 56: A phonetic transcription, spectrogram, and waveform for the utterance "will we ever forget it". Image credit to Alan Black.

17 Applications 2: Recognition/Generation of Continuous Inputs

While most of the previous sections have covered applications that take sequences of discrete inputs and generate sequences of discrete outputs, there are also a large number of works on modeling continuous inputs or ouputs, such as speech or images.

17.1 Automatic Speech Recognition

17.1.1 Characteristics of Speech and Speech Recognition

Speech is the method by which we communicate a large amount of the time, so assuming we are familiar with it on the high level. When processing speech by a computer, it is first input through a microphone as a waveform (bottom of Figure 56) corresponding to the change in air pressure. It can then be processed into a spectrogram, representing the strength at various frequency bands (middle of Figure 56) through the application of a Fourier transform. Segments in the speech correspond to **phonemes** such as "w" and "iy" corresponding to component sounds that make up words. Multiple phonemes together compose words, such as "w" and "iy" becoming "we".

One classical task that is a sequence-to-sequence modeling problem where the input sequence is continuous is **speech recognition** (often abbreviated ASR for "automatic speech recognition"). Speech recognition is difficult because the same word will never be said *exactly* the same way; the acoustic signal is filled with noise and/or speaker-specific characteristics and it is necessary to remove these texts. Speech recognition is generally evaluated using word error rate, which directly measures the number of insertions, deletions, or substitutions necessary to turn the output words into the reference text.

Modern speech recognition approaches can be generally split into two varieties: multi-component approaches that work by combining together multiple models, and end-to-end models that try to model speech recognition in a single large and directly-optimized model.

17.1.2 Multi-component Approaches

Approaches to speech recognition that use multiple components have at least two models: the *acoustic model* that makes a connection between the incoming acoustic signal, and a *language model* that scores the likelihood of the output. Usually there is also a *pronunciation dictionary* that maps sequences of phonemes into words, so the language model can be built on the word level. It is very common to formulate these models using weighted finite-state transducers (WFSTs), which will be described in Section 13 [10].

Acoustic models are now almost exclusively modeled using deep neural networks that either take in the acoustic features for a single frame x and predict its phoneme label y [9], or take in a whole sequence X, encode it with a network (such as bi-directional LSTMs [5] or Transformers [11, **?**]), and predict the probabilities based on this whole sequence worth of information. One interesting aspect of training acoustic models is that while we may know the speech signal and also the phonemes contained in the speech signal, we may not necessarily know the *alignment* between them, although we know that the speech signal and the phonemes must be in the same order. A method that can be applied to these problems **connectionist temporal classification** (CTC), which automatically induces an alignment between phonemes and corresponding frames using dynamic programming, and uses the alignments to train the neural network [4].

17.1.3 End-to-end Approaches

There have also been some promising preliminary results on end-to-end speech recognition with neural networks.

The most simple of them treats speech recognition as a regular sequence-to-sequence problem and solves it with encoder-decoder models [3]. In this case the encoder encodes the speech frames, and the decoder outputs the words or characters of the transcript. One difficulty is that sequences of frames included in a speech signal tend to be much longer than the sequences of words included in the corresponding transcript. Both for memory efficiency reasons, and to reduce the disconnect between the lengths of inputs and outputs, it is common to create architectures that reduce the length of the input sequence, including pyramidal RNNs [3], or strided CNNs [1] and Transformers [15].

There are also methods for speech recognition that directly try to predict words with CTC-based models [12]. However, in order to perform training efficiently using dynamic programming, CTC has to make an assumption of conditional independence of the output, and because of this it is common to incorporate a language model when actually generating results, moving these end-to-end models a bit closer to the component-based models above [21].

17.2 Speech Synthesis

Text-to-speech conversion, or **speech synthesis**, is the generation of speech from text, and models to do so generally stich together existing wave forms in a coherent way [6], or generate

speech using models such as hidden Markov models [20] and deep neural networks [19]. One method that has recently proven effective in the speech synthesis area uses dialated convolutional neural networks, which use convolutions with gradually increasing spans in the decoder portion of the network [17].[55] There are also methods for **voice conversion**, which map a sequence of speech frames to another sequence of speech frames in the voice of another speaker [16].

Speech synthesis models are often evaluated using **mel-frequency cepstral distortion** [8], which is a measure of difference between reference speech and the generated speech. This is an incomplete measure, however, and manual listening tests are often employed as well.

17.3 Speech Translation

The task of speech translation takes in speech and outputs either text or speech in another language. Again there are two approaches. The pipeline approach that first performs ASR to transform speech into text, translates the text from the source to target language, then optionally synthesizes target-language speech. In contrast, end-to-end speech translation attempts to go directly from source speech to target text or speech in a single model.

In the case of the pipeline approach, the simplest approach is relatively straightforward: you simply generate one-best results for each step in the pipeline. However, this can cause errors to propagate through each time step, leading to reduced accuracy. Thus, it can be useful to maintain ambiguity throughout the pipeline, for example by outputting n-best hypotheses or a graph-based structure encoding multiple ASR hypotheses, and feed this in to translation [13].

Recently, there have been impressive results in end-to-end speech translation, demonstrating that a model trained to go directly from speech to text in another language can do relatively similarly to a model using the gold-standard text on the source side, given sufficient training data [18]. One disadvantage of these methods is that they require a large corpus of aligned speech in the source language and text in the target language, which can often be a scarce resource compared to text-to-text translation data. To overcome this problem, there have been several methods that propose ways to additionally use additional data with intelligent multi-task learning strategies [2, 14].

17.4 Exercise

A potential exercise for this section would be to find and download a data set for one of these tasks, and run your sequence-to-sequence model on it and observe the results.

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[2] Antonios Anastasopoulos and David Chiang. Tied multitask learning for neural speech translation. *arXiv preprint arXiv:1802.06655*, 2018.

[55] These dilated convolutional networks have also proven useful in modeling text.[7]

[3] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[4] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. pages 369–376. ACM, 2006.

[5] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[6] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 373–376. IEEE, 1996.

[7] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

[8] John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68, 2008.

[9] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, 2012.

[10] Mehryar Mohri, Fernando Pereira, and Michael Riley. Speech recognition with weighted finite-state transducers. *Handbook on speech processing and speech communication, Part E: Speech recognition*, 2008.

[11] Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur. A time-restricted self-attention layer for asr. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE, 2018.

[12] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays. Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947*, 2015.

[13] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Neural lattice-to-sequence models for uncertain inputs. In *In Submission to Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[14] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Attention-passing models for robust and data-efficient end-to-end speech translation. In *Transactions of the Association of Computational Linguistics (TACL)*, 2019.

[15] Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. Self-attentional acoustic models. In *19th Annual Conference of the International Speech Communication Association (InterSpeech 2018)*, Hyderabad, India, September 2018.

[16] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

[17] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*, 2016.

[18] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly transcribe foreign speech. *arXiv preprint arXiv:1703.08581*, 2017.

[19] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7962–7966. IEEE, 2013.

[20] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication (SpeCom)*, 51(11):1039–1064, 2009.

[21] Thomas Zenkel, Ramon Sanabria, Florian Metze, Jan Niehues, Matthias Sperber, Sebastian Stüker, and Alex Waibel. Comparison of decoding strategies for ctc acoustic models. *arXiv preprint arXiv:1708.04469*, 2017.