LTI
Language, Technology and the Internet

Speech and Language Technology

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

Lecture 3

Revision of Representing Language

- Writing Systems
- Encodings
- Speech
- Bandwidth

Three Major Writing Systems

- Alphabetic (e.g., Latin)
  - one symbol for consonant or vowel
  - Typically 20-30 base symbols (1 byte)
- Syllabic (e.g., Hiragana)
  - one symbol for each syllable (consonant+vowel)
  - Typically 50-100 base symbols (1-2 bytes)
- Logographic (e.g., Hanzi)
  - pictographs, ideographs, sounds-meaning combinations
  - Typically 10,0000+ symbols (2-3 bytes)

Computational Encoding

- Need to map characters to bits
- More characters require more space
- Moving towards unicode for everything
- If you get the encoding wrong, it is gibberish

Speed is different for different modalities

Speed in words per minute (one word is 6 characters) (English, computer science students, various studies)

Modality	normal	peak
Reading	300	200 (proof reading)
Writing	31	21 (composing)
Speaking	150	
Hearing	150	210 (speeded up)
Typing	33	19 (composing)

- Reading >> Speaking/Hearing >> Typing

  $\Rightarrow$ Speech for input
  $\Rightarrow$ Text for output

Speech

- The need for speech representation
- Storing sound
- Transforming Speech
  - Automatic Speech Recognition (ASR): sounds to text
  - Text-to-Speech Synthesis (TTS): text to sound
- Speech technology — the Telephone!

The need for speech

- We want to be able to encode any spoken language
  - What if we want to work with an unwritten language?
  - What if we want to examine the way someone talks and don't have time to write it down?

- Many applications for encoding speech:
  - Building spoken dialogue systems, i.e. speak with a computer (and have it speak back).
  - Helping people sound like native speakers of a foreign language.
  - Helping speech pathologists diagnose problems

What does speech look like?

We can transcribe (write down) the speech into a phonetic alphabet.

- It is very expensive and time-consuming to have humans do all the transcription.
- To automatically transcribe, we need to know how to relate the audio signal to the individual sounds that we hear.
- We need to know:
  - some properties of speech
  - how to measure these speech properties
  - how these measurements correspond to sounds we hear

What makes representing speech hard?

- Sounds run together, and it's hard to tell where one sound ends and another begins.
- People say things differently from one another:
  - People have different dialects
  - People have different sized vocal tracts
- Hand-written text shares similar problems

- People say things differently across time: What we think of as one sound is not always (usually) said the same

- **coarticulation** = sounds affect the way neighboring sounds are said
  e.g. *k* is said differently depending on if it is followed by *ee* or by *oo*.

- What we think of as two sounds are not always all that different.
  e.g. The *s* in *see* is acoustically very similar to the *sh* in *shoe*

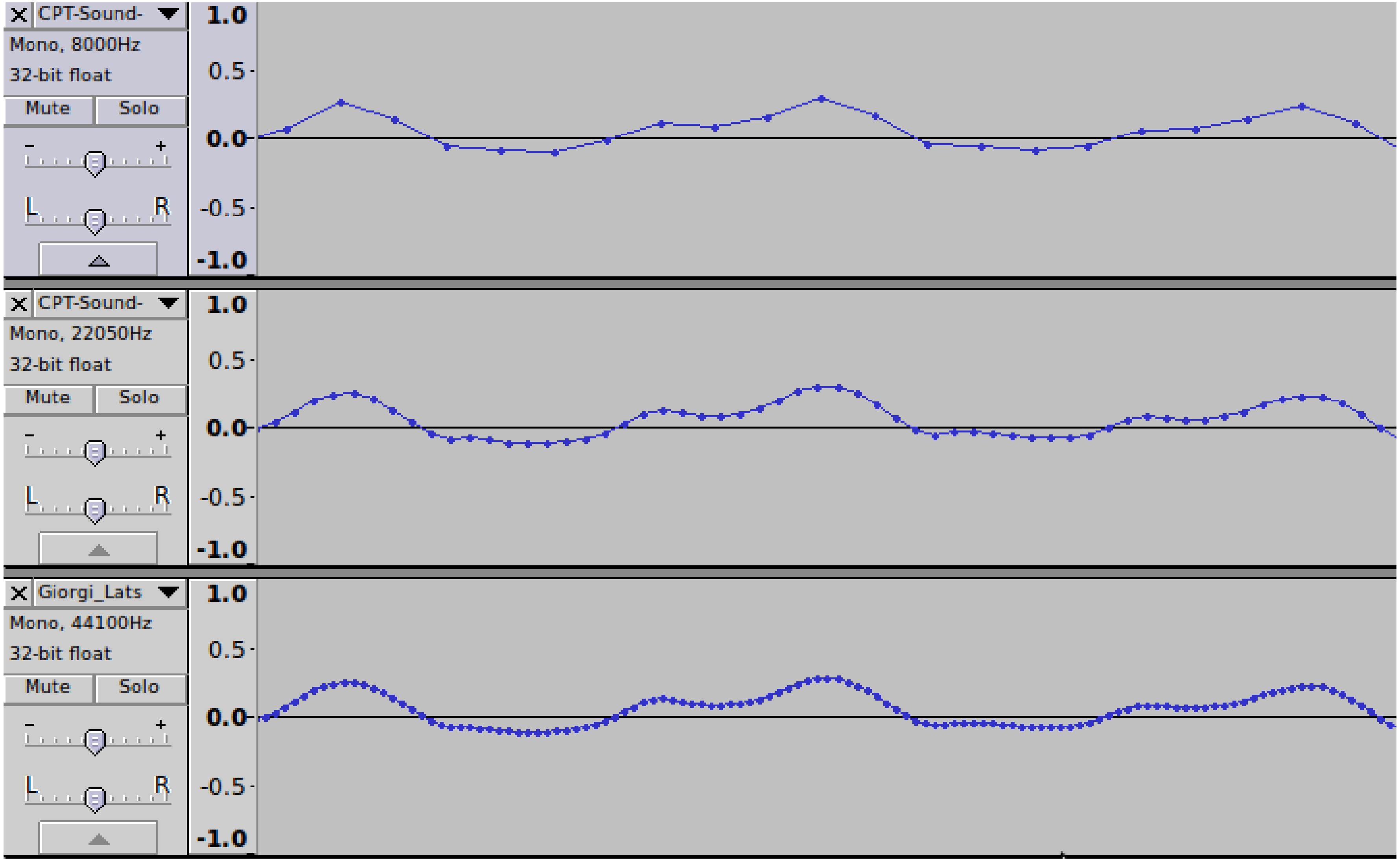
Articulatory properties: How it's produced

- We could talk about how sounds are produced in the vocal tract, i.e. articulatory phonetics
  - place of articulation (where): [t] vs. [k]
  - manner of articulation (how): [t] vs. [s]
  - voicing (vocal cord vibration): [t] vs. [d]
- But unless the computer is modeling a vocal tract, we need to know acoustic properties of speech which we can quantify.

Measuring sound

- Sound is actually a continuous wave
- We store data at each discrete point, in order to capture the general pattern of the sound
- Sampling Rate: how many times in a given second we extract a moment of sound; measured in samples per second
- Sound is continuous, but we prefer to store data in a discrete manner.

Signal sampling representation.



Comparison of a sound sample recorded at 8kHz, 22kHz and 44kHz.

Sampling rate

The higher the sampling rate, the better quality the recording ... but the more space it takes.

- Speech needs at least 8000 samples/second, but most likely 16,000 or 22,050 Hz will be used nowadays.
- The rate for CDs is 44,100 samples/second (or Hertz (Hz))

Now, we can talk about what we need to measure, ...

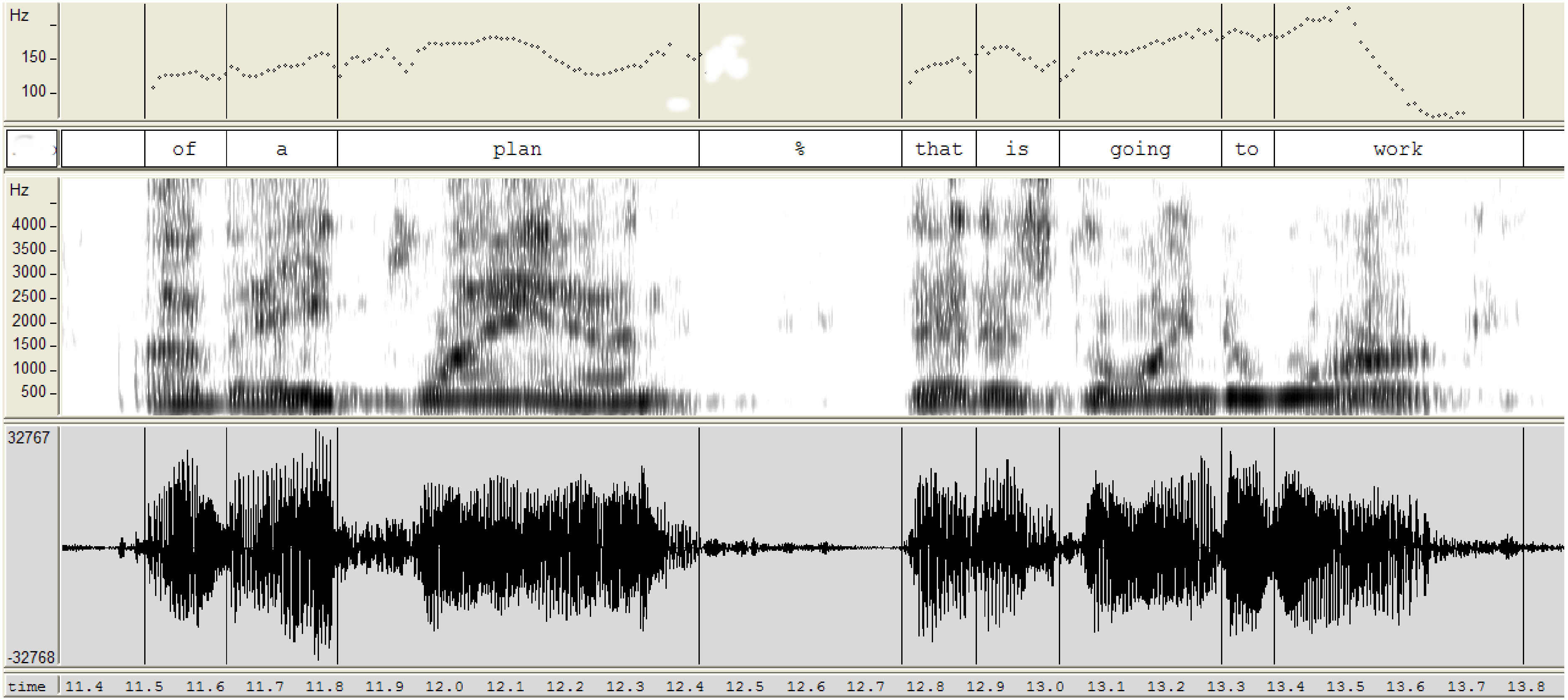
Acoustic properties: What it sounds like

- Sound waves: "small variations in air pressure that occur very rapidly one after another"

- The main properties we measure:
  - speech flow: rate of speaking, number and length of pauses (seconds)
  - amplitude (loudness): amount of energy (decibels)
  - frequency: how fast the sound waves are repeating (cycles per second, i.e. Hertz)
    - pitch: how high or low a sound is
    - In speech, there is a fundamental frequency, or pitch, along with higher-frequency overtones.

Researchers also look at things like intonation, i.e., the rise and fall in pitch

Speech Sample



Pitch track, transcription, spectogram and audio waveform.

Measurement-sound correspondence

- How dark is the picture? $\rightarrow$ How loud is the sound?
  - We measure this in decibels.
- Where are the lines the darkest? $\rightarrow$ Which frequencies are the loudest and most important?
  - We can measure this in terms of Hertz, and it tells us what the vowels are.
- Speech signals are very different from text.
  - No segmentation into words!

Applications of speech encoding

- Mapping sounds to symbols (alphabet), and vice versa, has some very practical uses.
  - Automatic Speech Recognition (ASR): sound to text
  - Text-to-Speech Synthesis (TTS): text to sound
- These are not easy tasks.
- Text-to-Speech Synthesis is somewhat easier.

Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR)

- Automatic speech recognition = process by which the computer maps a speech signal to text.
- Uses/Applications:
  - Dictation
  - Dialogue systems
  - Telephone conversations
  - People with disabilities –e.g. a person hard of hearing could use an ASR system to get the text (closed captioning)
  - Spying (many agencies run ASR on phone conversations and search for keywords)
  - Indexing audio data

Steps in an ASR system

1. Digital sampling of speech
2. Acoustic signal processing = converting the speech samples into particular measurable units
3. Recognition of sounds, groups of sounds, and words

May or may not use more sophisticated analysis of the utterance to help. e.g., a [t] might sound like a [d], and so word information might be needed (more on this later)

Kinds of ASR systems

Different kinds of systems, with an accuracy-robustness tradeoff:

- Speaker dependent: works for a single speaker
- Speaker independent: works for any speaker of a given variety of a language, e.g. American English
- A common type of system starts general, but learns
  - Speaker adaptive = start as independent but begin to adapt to a single speaker to improve accuracy
  - Adaptation may simply be identifying what type of speaker a person is and then using a model for that type of speaker
  - Or if it can get verification of it's hypothesis (e.g. did you click the search result), then it can add it as training data

Kinds of ASR systems

- Differing sizes and types of vocabularies
  - from tens of words to tens of thousands of words
  - normally very domain-specific, e.g., flight vocabulary

- continuous speech vs. isolated-word systems:
  - continuous speech systems = words connected together and not separated by pauses
  - isolated-word systems = single words recognized at a time, requiring pauses to be inserted between words
    - easier to find the endpoints of words
    - harder to use

Word Error Rate in Speech Recognition

- The first successful wide spread testing in NLP
  - Compare your output to a reference
  - Calculate the number of substitutions, deletions and insertions to make them match (Minimum Edit Distance)
  - Normalize by dividing by the length of the reference

$$WER = \frac{S+D+I}{N}$$

Reference:	I	want	to	recognize			speech	today
System:	I	want		wreck	a	nice	peach	today
Eval:			D	S	I	I	S	

- $WER = \frac{2+1+2}{6} = 0.83$

Some properties of WER

- Correlates well with the task
- Reducing WER is always a good thing
- A WER of 0 implies perfect results
  (assuming the reference is correct)
- $WER < .05$ considered the minimum to be useful
- Competitions were held to see who could get the lowest WER
  - Speech Recognition had 10 years of rapid improvement
  - It has slowed down now

How good are the systems?

Task	Vocab	WER (%)	WER (%) adapted
Digits	11	0.4	0.2
Dialogue (travel)	21,000	10.9	—
Dictation (WSJ)	5,000	3.9	3.0
Dictation (WSJ)	20,000	10.0	8.6
Dialogue (noisy, army)	3,000	42.2	31.0
Phone Conversations	4,000	41.9	31.0

Results of various DARPA competitions (from Richard Sproat's slides, 2012)

Improvements in machine learning (**deep learning**) have further reduced errors

- A combination of learning a combined model and better training data
  Improving End-to-End Models For Speech Recognition (Google AI Blog 2017)
  WER of 5.6% (16% relative improvement over 6.7%)
  - Teaching the Google Assistant to be Multilingual (2018)
  - Looking to Listen: Audio-Visual Speech Separation (2018)

How good are the systems in 2020?

English Tasks	WER%
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3

Chinese (Mandarin) Tasks	CER%
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

Rough Word Error Rates (WER = % of words misrecognized) reported around 2020 for ASR on various American English and Chinese recognition tasks. Note, the error rate for transcribing conversations between humans is much higher; 5.8 to 11% for the Switchboard and CALLHOME corpora.

From Daniel Juravsky and James H. Martin (2023) Chapter 16 Automatic Speech

Recognition and Text-to-Speech in *Speech and Language Processing*. Draft of January 7, 2023.

How good are the systems in 2024?

- For well-resourced languages, very, very good.
- I have used Whisper (from openai) with great success on
  - Comedy show (recorded on phone) - English
  - Interview with informants (on phone, in noisy restaurant, on boats, ...) - English and Spanish
  - Interview in quiet room (on phone) - Mandarin

  All were generally astoundingly accurate
  with random places where the model completely failed!

- Large language models are good but inpredictable

Why is it so difficult?

- Speaker variability
  - Gender
  - Dialect/Foreign Accent
  - Individual Differences: Physical differences; Language differences (idiolect)
- Many, many rare events
  - 300 out of 2,000 diphones in the core set for the AT&T NextGen system occur only once in a 2-hour speech database

Rare events are frequent

- Collect about 10,000,000 character 4-grams, from English newswire text, merging upper and lower case —60 distinct characters including space.
- 197,214 lines of text.
- Of these, 14,317 (7%) contain at least one 4-gram that only occurs once in 10,000,000.
- Increase it to 5-grams: 21% of lines contain contain at least one 5-gram that only occurs once in 10,000,000.

What is an n-gram?

- An n-gram is chunk of n things: most often words, but could be characters, letters, morphemes, stems, ...
- Approximation of language: information in n-grams tells us something about language, but doesn't capture the structure
- Efficient: finding and using every, e.g., two-word collocation in a text is quick and easy to do
- n-grams help a variety of NLP applications, including word prediction
  - We can predict the next word of an utterance, based on the previous
- *unigram, bigram, trigram, 4-gram, ...*

Mozilla Common Voice

- a crowdsourcing project to create a free database for speech recognition software
- volunteers record sample sentences with a microphone and review recordings of other users
- transcribed sentences are collected in a voice database available under the public domain license CC0
- In 2020, there were 40 languages, with 3401 validated hours
- a good example of citizen science (or engineering)

Systems improve with more data or better algorithms, we need work on both.

Text-to-Speech Synthesis (TTS)

Text-to-Speech Synthesis (TTS)

- Could just record a voice saying phrases or words and then play back those words in the appropriate order.

- This won't work for, e.g., dialogue systems where speech is generated on the fly.

- Or can break the text down into smaller units

  1. Convert input text into phonetic alphabet (ambiguous mapping)
  2. Synthesize phonetic characters into speech

- To synthesize characters into speech, people have tried:

  - using a model based on frequencies, the loudness, etc.
  - using a model of the vocal tract and human speech production

Demo of Festival

Festival – a current system:
http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

HTS - a statistical parametric approach (both the 2005 and 2007 systems)

Unit - standard unit selection concatenative approach
look for variable-length units in an annotated database of speech, and select them on the basis of various features including desired phoneme sequence and prosody. Units can be individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences.

Diphone - single instance diphone concatenation
(the previous TTS generation technology, from mid 1980's to mid 1990's).

Two steps in a TTS system

1. Linguistic Analysis

   - Sentence Segmentation
   - Abbreviations: *Dr Smith lives on Nanyang Dr. She is ...*
   - Word Segmentation:
     - 森山前日銀総裁 *Moriyama zen Nichigin Sousai*
     - ⊗ 森山前日銀総裁 *Moriyama zennichi gin Sousai*

2. Speech Synthesis

   - Find the pronunciation
   - Generate sounds
   - Add intonation

Linguistic Analysis (cont)

- Acronyms: *NTU, NATO*
- Numbers: *666 green bottles; They were branded with 666.*
- Senses: *Star Wars IV; IV drip* (“four vs “intravenous”)
  Are you content with the content?
  The bandage was wound round the wound.
  Polish polish should be used.
- Inflection:

  statement falling intonation
  question rising intonation
  ...

Segmental durations:

- Every sound has to have some time assigned to it
- Other things being equal:
  - Vowels tend to be longer than consonants
  - Stressed segments tend to be longer than unstressed segments
  - Accented segments tend to be longer than unaccented segments
  - Final segments tend to be longer than non-final segments
  - Segments have different inherent durations:
    /ee/ in *keep* is generally longer than /i/ in *kip*

Synthesizing Speech: Analysis

- From linguistic analysis we have:
  - A set of sounds to be produced
  - Associated durations
  - Associated fundamental frequency information
  - Possibly other things:
    - Amplitude
    - Properties of the vocal production
- Now we are ready to synthesize speech

Speech Synthesis

Almost all speech synthesis is now done using deep-learning where neural networks are trained on recorded speech paired with the input text.

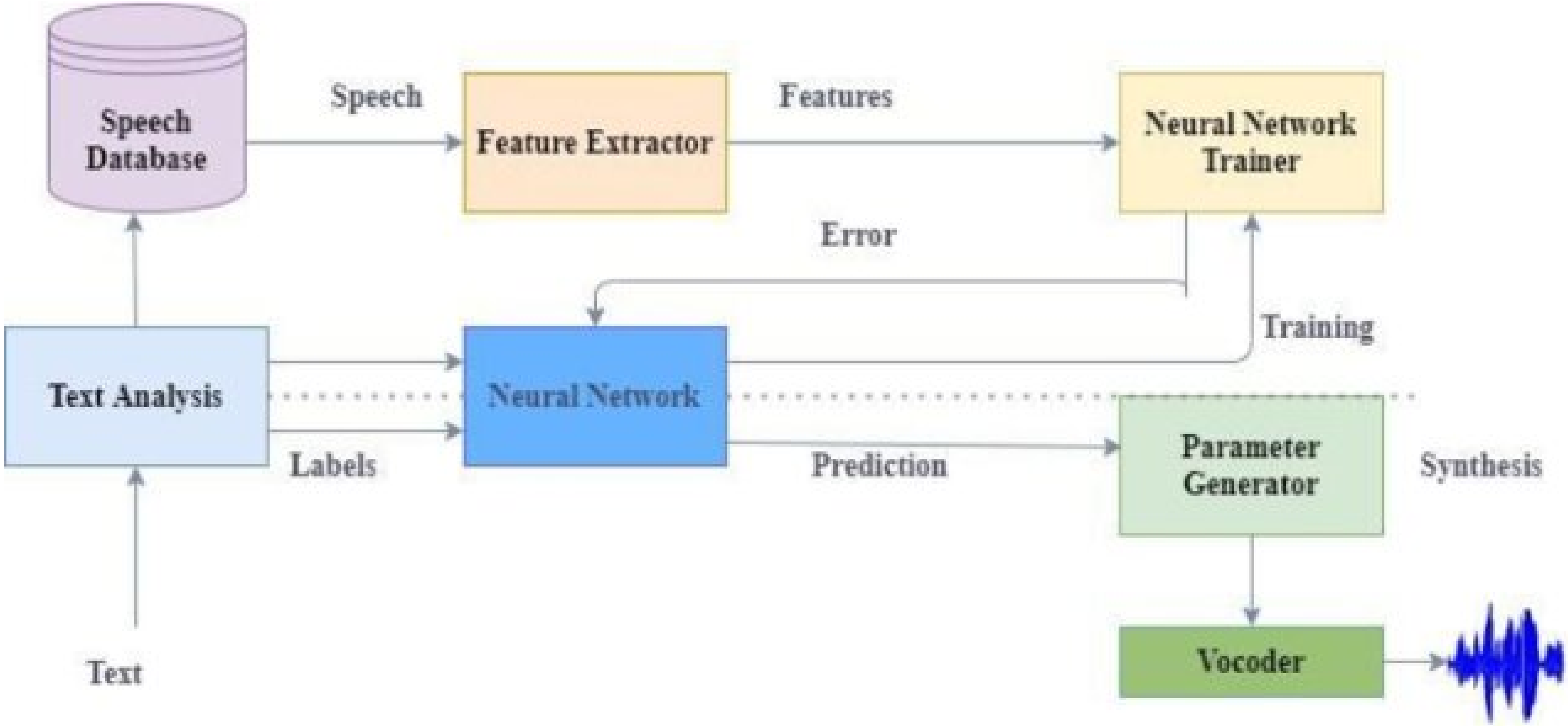


Image from (Khanam et al., 2022, Figure 7)

Speech Synthesis with Neural Networks

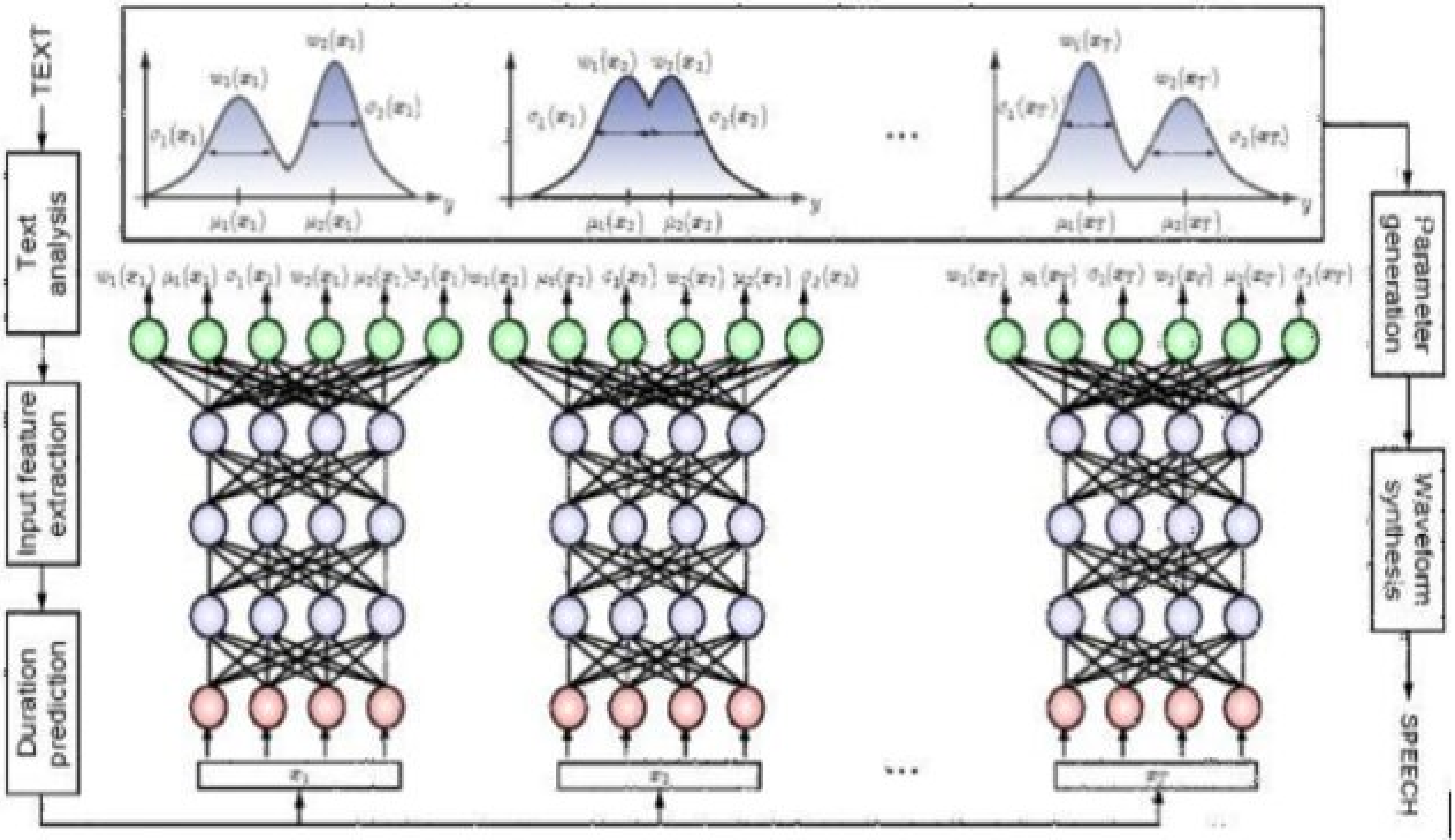


Image from (Khanam et al., 2022, Figure 6)

Prosody of Emotion

- Excitement: Fast, very high pitch, loud
- Hot anger: Fast, high pitch, strong, falling accent, loud
- Fear: Jitter
- Sarcasm: Prolonged accent, late peak
- Sad: Slow, low pitch

The main determinant of “naturalness” in speech synthesis is not “voice quality”, but natural-sounding prosody (intonation and duration)

Richard Sproat

It's hard to be natural

When trying to make synthesized speech sound natural, we encounter the same problems that make speech encoding hard:

- The same sound is said differently in different contexts.
- Different sounds are sometimes said nearly the same.
- Different sentences have different intonation patterns.
- Lengths of words vary depending on where in the sentence they are spoken.
  1. The car crashed into the tree.
  2. It's my car.
  3. Cars, trucks, and bikes are vehicles.

Speech to Text to Speech

If we convert speech to text and then back to speech, it should sound the same.

- But at the conversion stages, there is information loss.
- To avoid this loss would require a lot of memory and knowledge about what exact information to store.
- The process is thus irreversible.
- In fact, people can't say the same sentence exactly the same way either!

TTS Applications

Any situation where you need information, but can't access it visually:

- Access to information for the blind
- Access to email, news, stock quotes ...over the phone
- Directions to drivers
- Spoken dialog systems where it is not practical to prerecord everything
- Informational content –e.g. NOAA Weather Radio –where it would be expensive to have a human read all the announcements.
- This has also increased in quality with deep learning
- Transforming speech is increasingly easy — from one voice to another, ...
  - Get the prosody from a human, match to your needs

Mediums of Communication

Mediums of Communication

- Different mediums of communication
  - affect the language used within them
  - may affect our social organization
- We will analyze them compared to speech/text
  - More fine grained analyses exist (Herring, 2007)

The Telephone

Speech like	Text like
time bound	space bound
spontaneous	contrived
face-to-face	visually decontextualized
loosely structured	elaborately structured
socially interactive	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

- Technology enabling a new modality of communication
- Speech-like but not exactly speech
- Analysis from Crystal (2006)

Phone Schema

1. Greeting/Introduction
   Hello. This is ∼. Thank you for calling ∼.
   jpn: *moshi-moshi*; **kor**: *yeobo seyo*

2. Connecting: *May I speak to ∼. I'll put you through.*

3. Meta-requests
   Can you call me back? I think we have a bad connection.
   Can you please hold for a minute? I have another call.

4. Taking a message
   Can I ask who's calling? Would you like to leave a message?

5. Finishing: *Thanks for calling. Bye for now.*

Conventions for dealing with the new technology

Phone Greetings in Different Langauges

- ITALIAN
  - In Italy, the common greeting is *Pronto*. That translates roughly to "Ready," as in, "I'm here and can hear you."
- POLISH
  - The Polish greeting is *Tak. Słucham?*. The question being asked: "Hello, who is it calling?"
- SPANISH
  - In some Spanish-speaking countries, you'd say *¿Diga?* That means "speak," or "you can go ahead and start talking now."
- SPANISH in MEXICO
  - On the phone, you'd say *bueno*. That literally means "good" in English, but in this context it means something more like "well?"

Taken from: https://www.portablepress.com/blog/2017/05/moshi-moshi/

Effects of the telephone

- The telephone (and telegraph) had a big effect on independence of subsidiaries in large international organizations (Parkinson, 1958)
  - Central offices could micromanage people in the field
  - More centralization, less local flexibility

Acknowledgments and References

- Many slides on speech technology adapted from Richard Sproat's L270: http://catarina.csee.ogi.edu/L270/

- Crystal, D. (2006). *Language and the Internet*. Cambridge University Press, 2nd edition
- Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. Language@Internet. http://www.languageatinternet.org/articles/2007/761
- Parkinson, C. N. (1958). *Parkinson's Law, or The Pursuit of Progress*. John Murray, London
- Speech synthesis diagrams from: Khanam, F., Munmun, F., Ritu, N., Saha, A., and Mridha, F. (2022). Text to speech synthesis: A systematic review, deep learning based architecture and future research direction. *Journal of Advances in Information Technology*, 13