Lecture 10: Automatic Recognition of 200 Words
Velichko & Zagoruyko, 1970

Mark Hasegawa-Johnson

ECE 537: Speech Processing Fundamentals

Outline

Automatic Speech Recognition (ASR)

- Control sequence (cs): a sequence of 203 spoken words that you want to recognize
- Training sequence (ts): a second recording of each of those 203 words

1. **один**—od'ín—one
2. **два**—dvá—two
3. **три**—tri'í—three
4. **четыре**—tʃetir'e—four
5. **пять**—pját—five
6. **шесть**—ʃést'—six
7. **семь**—s'ém'—seven
8. **восемь**—vós'em'—eight
9. **девять**—d'évjat'—nine
10. **ноль**—nól—zero
11. **плюс**—pljús—plus
12. **минус**—m'ínus—minus
13. **разделить**—razd'el'ít'—divide

Automatic Speech Recognition

"Automatic speech recognition" (ASR) means that, for each word in cs, find the word in ts that is most acoustically similar.

- If it's the same word, "correct"
- Otherwise, "error"

1. один—od’ín—one
2. два—dvá—two
3. три—tri’í—three
4. четыре—tʃetir’e—four
5. пять—pját—five
6. шесть—ʃést’—six
7. семь—s’ém’—seven
8. восемь—vós’em’—eight
9. девять—d’évjat’—nine
10. ноль—nól—zero
11. плюс—pljús—plus
12. минус—m’ínus—minus
13. разделить—razd’el’ít’—divide

Error Rate

- For each of 4 different ts,
  - for each of 3 different cs,
    - Compute # correct out of 203 words in the cs
  - Recognition reliability for the first ts is

$$\frac{609-26}{609}=0.957$$

TABLE 2

Recognition results of speaker No. 2

N ts \ N cs	No. of errors 1	2	3	4	Recognition reliability
1		10	7	9	95·7
2	12		8	10	95
3	5	6		9	96·7
4	3	9	8		96·7

cs, Control sequence; ts, training sequence.

What makes two words similar?

- This method demands the following question: how do we measure the acoustic similarity between two recorded words?
- Answer: dynamic time warping, using log-spectral features.

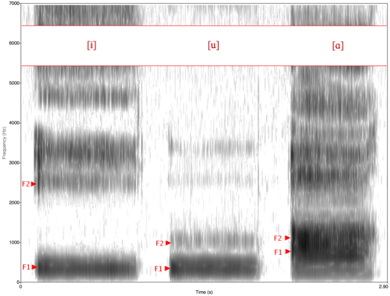
Outline

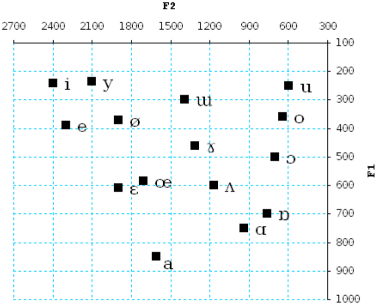
Log-Spectral Features for the 200-Word Speech Recognizer

- Spectral features included the log energy in five frequency bands.
- Constant-Q filters are motivated by auditory processing.
- Logarithmic units are motivated by the Weber-Fechner law.
- Euclidean distance between log-energy spectra is inverted to compute similarity.

Center Frequencies: voiced, high, back, front, fricated

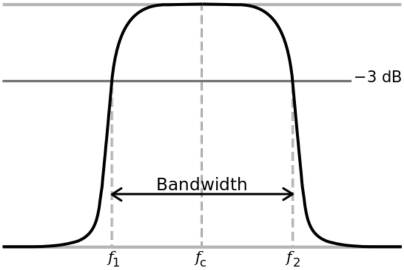
Five bandpass-filtered signals are computed, w/center frequencies 225, 450, 900, 1800, 7200Hz. These correspond roughly to measurements of voicing, tongue height, tongue backness, tongue frontness, and frication.

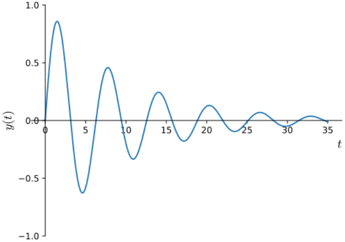




Auditory filters tend to have higher bandwidth at higher frequencies. V & Z model this phenomenon using a constant-Q analysis, with $Q = 2.45$. Quality of a filter (Q) is center freq over bandwidth, $Q = \frac{f_c}{B}$. It is also the number of undamped oscillation periods of the impulse response:

$$h(t) = e^{-\pi B t} \sin(2\pi f_c t) u(t)$$





Constant-Q Analysis

Using constant $Q = 2.45$, we get the following bandwidths for the V& Z sub-bands:

Center Frequency f_c (Hertz)	Bandwidth $B = \frac{f_c}{2.45}$ (Hertz)
225	92
450	184
900	367
1800	735
7200	2939

Bandpass Energies

- V& Z computed bandpass filters in continuous time, but let's pretend discrete time: $x_i[n] = h_i[n] * x[n]$.
- The sub-band energy is the squared signal, summed over one frame:

$$E_i = \sum_{n=0}^{N-1} (x_i[n])^2$$

- The signal energy is

$$E_0 = \sum_{n=0}^{N-1} (x[n])^2$$

- V& Z use the following features, which are guaranteed to be non-negative:

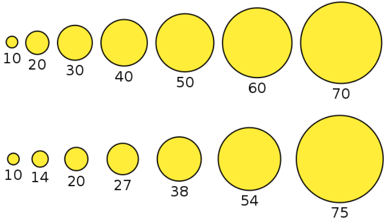
$$f_i = \ln\left(\frac{E_0}{E_i}\right)$$

Weber-Fechner Law

- Features are $\ln E_0/E_i$. Logarithm is motivated by the Weber-Fechner Law.
- The Weber-Fechner law says that the minimum noticeable increase ΔI of intensity for a sense organ is proportional to intensity itself I:

$$\frac{\Delta I}{I} = \text{constant}$$

- If loudness followed the Weber-Fechner law, it would be measured by decibels.



Spectral Similarity

Suppose we have two speech segments characterized by the spectral features $\ln\left(\frac{E_0^{(i)}}{E_d^{(i)}}\right)$ for segment i, and $\ln\left(\frac{E_0^{(k)}}{E_d^{(k)}}\right)$ for segment k. Calculate the Euclidean distance between these two spectra:

$$\rho_{i,k} = \sqrt{\sum_{d=1}^{5}\left(\ln\left(\frac{E_0^{(i)}}{E_d^{(i)}}\right) - \ln\left(\frac{E_0^{(k)}}{E_d^{(k)}}\right)\right)^2}$$

"Similarity" is the regularized inverse of distance:

$$a_{i,k} = \frac{2}{2+\rho_{i,k}^2}$$

Outline

1. Automatic Speech Recognition
2. Log-Spectral Features
3. Dynamic Time Warping
4. Conclusion

How to Measure Similarity Between Two Spoken Words

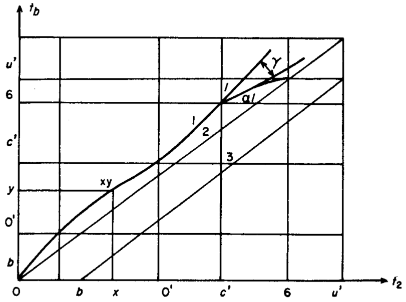
- Call the shorter word the “vertical” word. It is a sequence of m frames, $1 \leq i \leq m$ (each frame is a five-dimensional log spectrum).
- The longer word is the “horizontal” word. It is a sequence of n frames, $1 \leq k \leq n$, $n \geq m$.
- The similarity between frame i and frame k is $a_{i,k}$.

How to Measure Similarity Between Two Spoken Words

Linear Time Warping computes word similarity by stretching one word to match the other, then averaging the frame similarities:

$$B = \frac{1}{m}\sum_{i=1}^{m} a_{i,k=\left(\frac{n}{m}\right)i}$$

This is shown as line 2 in the figure.

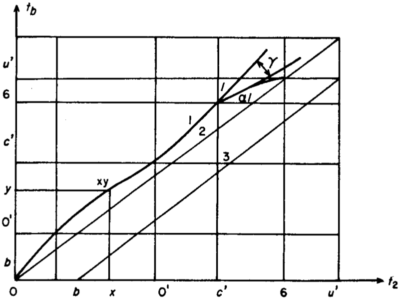


How to Measure Similarity Between Two Spoken Words

Linear Time Warping with Shift computes word similarity on a straight line with a shift:

$$B = \frac{1}{m} \sum_{i=0}^{m(1-b/n)} a_{i,k=\left(\frac{n}{m}\right)i+b}$$

This is line 3 in the figure.

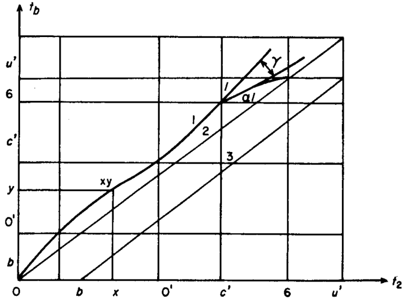


Dynamic Time Warping

Dynamic Time Warping computes word similarity by finding the alignment curve that maximizes B:

$$B = \frac{1}{m} \max_{k(1),\ldots,k(m)} \sum_{i=1}^{m} a_{i,k(i)}$$

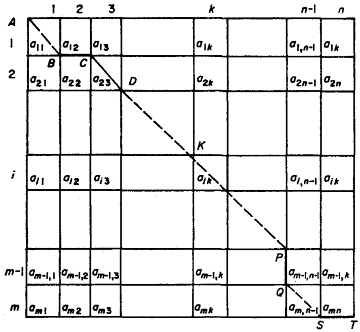
...subject to the constraint that neither time axis ever goes backward ($-\frac{\pi}{4} \leq \gamma \leq \frac{\pi}{4}$). This is curve 1 in the figure.



Dynamic Time Warping

The curve of maximum similarity can be computed by dynamic programming:

- **Initialize:** $A_{m+1,k} = A_{i,n+1} = 0$ for all i, k.
- **Iterate:** $A_{i,k} = \max\left(A_{i+1,k}, A_{i,k+1}, a_{i,k} + A_{i+1,k+1}\right)$
- **Terminate:** $B = \frac{1}{m} A_{1,1}$.



Insertions, Deletions, and Substitutions

$$A_{i,k} = \max\left(A_{i+1,k}, A_{i,k+1}, a_{i,k} + A_{i+1,k+1}\right)$$

Notice there are three possible step directions:

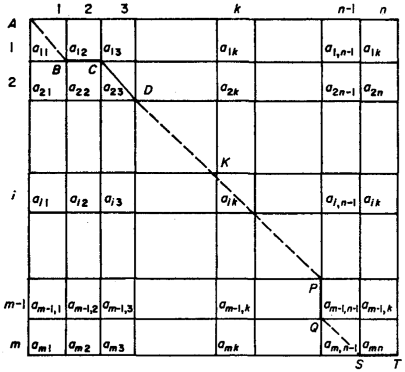
- Vertical: $A_{i,k} = A_{i+1,k}$, frame i is inserted.
- Horizontal: $A_{i,k} = A_{i,k+1}$, frame k is deleted.
- Diagonal: $A_{i,k} = a_{i,k} + A_{i+1,k+1}$, frame i is substituted for frame k.

The algorithm chooses as many diagonal steps as it can, because $a_{i,k} \geq 0$.

Insertions, Deletions, and Substitutions

- The algorithm chooses as many diagonal steps as it can, because $a_{i,k} \geq 0$.
- The largest possible number of diagonal steps is m.
- Therefore, I think the the average per-frame similarity should be normalized by $\frac{1}{m}$; I think the $\frac{1}{n}$ in the article is a typo, but I'm not sure!

$$B = \frac{1}{m} A_{1,1}$$



Computational Complexity

- Linear time warping is $\mathcal{O}\{n\}$ per word-pair, because it only tests one alignment.
- Dynamic time warping is $\mathcal{O}\{n^2\}$ per word-pair, to test every alignment.
- If there are v words in the training sequence, complexity is $\mathcal{O}\{n^2 v\}$ per test word.
- Z& V reduce complexity by using the following algorithm. For each test word,
  1. Use LTW for all training words, choose 32.
  2. Use LTW+shift with s different shifts, choose 8 best words.
  3. Use DTW to find the 1 best.

  Total complexity: $\mathcal{O}\{8n^2 + 32sn + vn\}$ per test word.

Outline

Summary

- Similarity of two words is defined to be the maximum, among all possible alignments, of the average similarity of the aligned spectra.
- This is computed by dynamic programming (DP):

$$A_{i,k} = \max\left(A_{i+1,k}, A_{i,k+1}, a_{i,k} + A_{i+1,k+1}\right)$$

- Similarity of any pair of spectra is $a_{i,k} = \frac{2}{2+\rho_{i,k}^2}$,

$$\rho_{i,k} = \sqrt{\sum_{d=1}^{5}\left(\ln\left(\frac{E_0^{(i)}}{E_d^{(i)}}\right) - \ln\left(\frac{E_0^{(k)}}{E_d^{(k)}}\right)\right)^2}$$