Speech and Audio Processing (Speech)

EXAM

Spring 2020

---

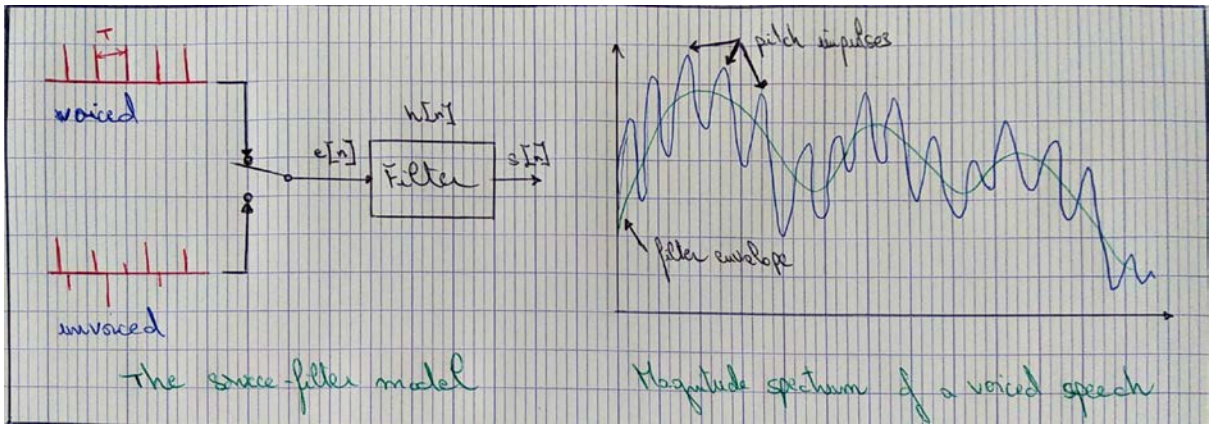
1. Explain the difference between deterministic and stochastic approaches to the automatic recognition of speech signals and give examples of each.

Deterministic approaches are based on a straightforward comparison between signal instances (e.g. time-series) by extracting specific properties. **D**eterministic **T**ime **W**arping is an example of deterministic approaches. It consists of finding similarities between two time-series by defining a metric to calculate "distance" between instances.

Stochastic approaches involve the use of probabilistic models based on the statistical properties of the signal. Since different acoustic performances are never the same, his approach seems to be more adapted for speech recognition because it takes into consideration speech uncertainty caused by speaking rate, variability, etc. The **H**idden **M**arkov **M**odel is an example of this approach.

2. Sketch the source-filter model and explain the parts or components of the human speech production mechanism that influence the source and the filter characteristics. Sketch an example magnitude spectrum of a voiced speech segment derived with the discrete Fourier Transform and explain which properties correspond to the source and which correspond to the filter.

The source-filter model represents speech as a result of a linear filter modeling the resonant characteristics of the vocal tract, which is applied to the source signal modeling the glottal excitation (release of pressure behind the glottis). In the time domain, it's represented as a convolution between source and filter, i.e. $s(t) = e(t) * h(t)$, which is simply a multiplication



in the frequency domain, i.e. $S(\omega) = E(\omega) \times H(\omega)$. The source-filter model is described in the next sketch.

The first thing to mention is that the voiced pitch is characterized by its high energy. The excitation, which is an impulse train, is multiplied by the filter which is an envelope with visible formants. The result is the pitch structure represented previously.

3. Explain why hidden Markov models are referred to as hidden and why they are a form of generative classifier.

HMMs are referred to as hidden because we cannot observe the state sequence from the observation (but we still can infer them). In other words, the observed symbols do not uniquely define a state.

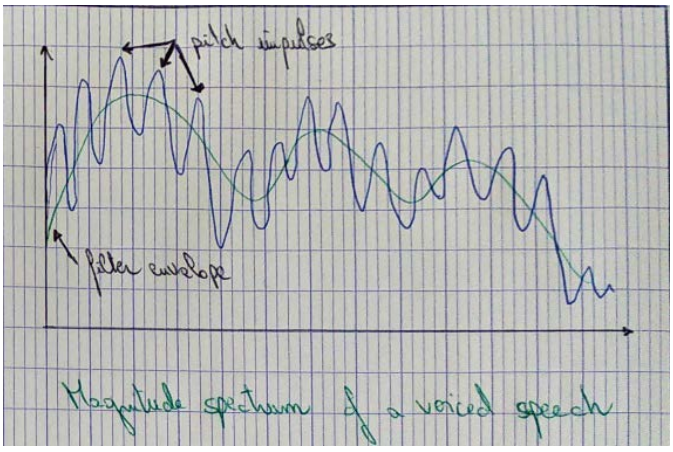
HMMs are generative models, which means we are interested in the probability that this model could have generated the sequence observation we have.

4. Explain why speech signals are usually processed in small segments called frames, rather than being processed in one segment in full.

Speech is a non-stationary signal but it is assumed to be stationary during short intervals of a typical length of about $10ms$. Stationary means that during these intervals the frequency or spectral characteristics are not changing. That's why it's processed in frames.

5. What accounts for the pitch and formants of a speech signal? Sketch an example magnitude spectrum for a voiced speech segment and describe the aspects of it that correspond to the pitch and formants.

The voiced speech is generated by the vibration of the vocal cords located in the glottis opening and closing at a specific frequency.
It is characterized by its periodicity in the time domain. Its spectrum is featured as some peaks with a formant envelope. The peaks represent the pitch period and the formants reflect the vocal tract feature.

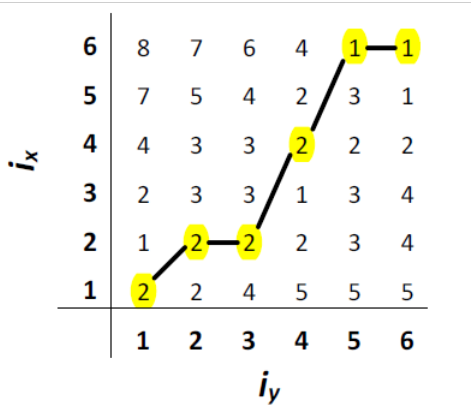


6. Explain the role and importance of the inverse Discrete Fourier transform (IDFT) in the extraction of basic cepstral features and the discrete cosine transform (DCT) in the extraction of Mel-scaled cepstral features. Comment briefly on the differences between the two.

The inverse Discrete Fourier transform is the last operation, applied on the logarithm of the magnitude of the DTFT of the signal, to get the Cepstrum which is defined as $c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{i\omega})|\, e^{i\omega n} d\omega = \mathcal{F}^{-1}(\log|\mathcal{F}(x[n])|)$, where $X(e^{i\omega})$ is the DTFT of the signal.
The IDFT is used to transform the log-magnitude of the signal (after liftering) to obtain the filter properties, i.e. $h[n]$.

The Discrete Cosine Transform is applied after getting the log-energy coefficients within the filterbank. This is a simplified version of FT because it results in real-value and decorrelated coefficients, so we can get a more compact representation (e.g. diagonal correlation matrix).

7. Using the local path constraints and slope weights shown, demonstrate the use of the dynamic time warping (DTW) algorithm in finding the best warping path in the following grid where each point and number within the grid indicates the distance between reference and test frames and i_x and i_y indicate the frame indices.



The best wrapping path respecting the local constraints is represented in the previous figure. The total distance between the two time-series instances is 15 (one to one distance plus transitions).

8. Stating any assumptions that you make, design an automatic speech recognition system to distinguish between a small set of words. It should be based upon two-state, left-to-right, word-level hidden Markov models (HMMs) and cepstral feature representations that are vector quantized into three observation

symbols K, L and M. Using the initial model parameters below, learn a model $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ (where each symbol has its usual meaning) for one of the words, which has training sequence observations KKLMM.

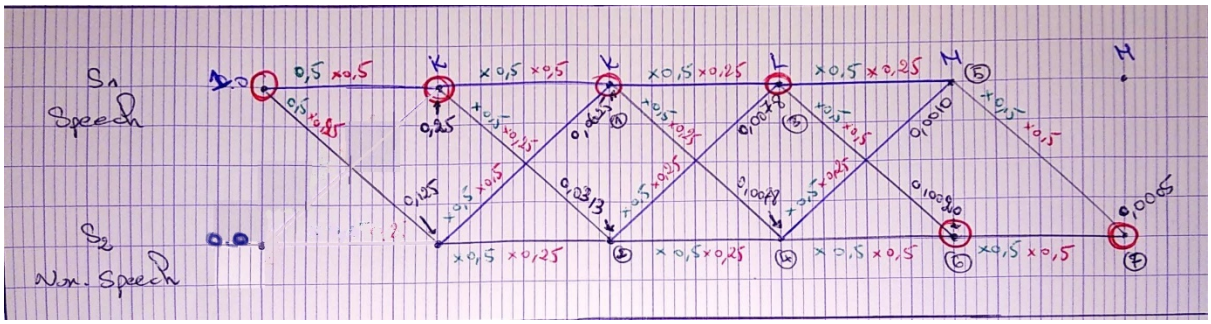
$$\boldsymbol{\pi} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{A} = \begin{bmatrix} \mathbf{0.5} & \mathbf{0.5} \\ \mathbf{0.5} & \mathbf{0.5} \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \mathbf{0.5} & \mathbf{0.25} \\ \mathbf{0.25} & \mathbf{0.25} \\ \mathbf{0.25} & \mathbf{0.5} \end{bmatrix}$$

Imagine that you train models for the full set of words and then run a set of experiments from which you find performance to be poor. Describe some steps that you could take to improve performance.

Since we are working on a word-level hidden Markov model, we can make a few assumptions:

- The two states correspond to speech (S_1) and non-speech (S_2).
- Every word must end with a non-speech state (S_2).

The first step is to determine the most likely state sequence using the Viterbi algorithm.



The likelihood of each state at every t is calculated as described next.

1. $\max(0.25 \times 0.5 \times 0.5, 0.125 \times 0.5 \times 0.5) = 0.0625$
2. $\max(0.25 \times 0.5 \times 0.25, 0.125 \times 0.5 \times 0.25) = 0.03125$
3. $\max(0.0625 \times 0.5 \times 0.25, 0.0313 \times 0.5 \times 0.25) = 0.0078$
4. $\max(0.0625 \times 0.5 \times 0.25, 0.0313 \times 0.5 \times 0.25) = 0.0078$
5. $\max(0.0078 \times 0.5 \times 0.25, 0.0078 \times 0.5 \times 0.25) = 0.0010$
6. $\max(0.0078 \times 0.5 \times 0.5, 0.0078 \times 0.5 \times 0.5) = 0.0020$
7. $\max(0.0010 \times 0.5 \times 0.5, 0.0020 \times 0.5 \times 0.5) = 0.0005$

So we can identify the most likely state sequence $S_1S_1S_1S_1S_2S_2$ and then update the model parameters (we can also choose $S_1S_1S_1S_2S_2S_2$ which has the same likelihood).

	S_1	S_2
K	2	0
L	1	0
M	0	2
Σ	3	2

		from	
		S_1	S_2
to	S_1	2	1
	S_2	1	1
	Σ	3	2

$$\bar{\pi} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \bar{A} = \begin{bmatrix} 0.66 & 0.5 \\ 0.33 & 0.5 \end{bmatrix}, \bar{B} = \begin{bmatrix} 0.66 & 0 \\ 0.33 & 0 \\ 0 & 1 \end{bmatrix}$$

To improve performance, we can use other instances to train every word's model, and update the parameters again.

[END OF PAPER]