EECS 225d: Midterm 2, April 25, 1997

Version with correct answers (not necessarily unique)

Name:

For multiple choice questions, also write a brief (1-2 sentence) explanation of why the answer is correct. The parenthesized numbers after each question number give the relative point value of each question.

1. (2) In cepstral analysis of speech, the vocal tract characteristics would most clearly be seen in

   a) the zeroth coefficient
   b) low-time coefficients
   c) high-time coefficients
   d) equally in all coefficients

   ⇒The formant (resonant) structure of the vocal tract frequency response primarily affects the low-time ceptral coefficients (b); for voiced sounds, the higher coefficients are more indicative of the pitch.

2. (2) LPC spectral analysis

   a) yields a pole-only representation
   b) can be more affected by pitch for children than for adults
   c) tends to model spectral peaks better than valleys
   d) all of the above

   ⇒LPC analysis yields a pole-only representation, since it corresponds to an autoregressive model of speech production; however, the other choices above are also true. Voiced sounds with high pitch (such as those produced by children) have relatively few harmonic components, and so the spectral estimate can be strongly effected by the fundamental frequency. Finally, due to the pole-only representation and the squared error criterion, peaks are represented more faithfully than valleys. Therefore, (d) is the correct answer.

3. (2) A local slope constraint for DTW could be

   a) The start of an input word cannot be matched to the end of a reference template

   b) Subtract the local distance from the best global distance

   c) Cross-country skis are not permitted for the higher slopes

   d) Predecessor global distances used to compute the current global distance are constrained to the previous or current input or reference frames.

   ⇒Selection (d) corresponds to the local slope constraint - that is, for each local decision in the dynamic programming, there is a constraint on legal predecessors. While selection (a) is a good constraint, it is a global one.

4. (2) A difference between a Hidden Markov Model (HMM) and a Markov Model (MM) is

   a) HMMs have better hiding places.

   b) MMs don't have observations.

   c) MM observations are deterministic, given a state transition

   d) HMM states have stochastic transitions

   ⇒Selection (c) is correct - for HMMs, given a state transition we only know a density function for observations, while in the MM case the observation is determined by the transition.

5. (2) Mel cepstra, PLP cepstra, and LPC cepstra are all sometimes used for the front end of ASR systems. A major characteristic of all three analysis methods (in contrast to power spectra estimated as the squared magnitude of the DFT of the windowed data) is

   a) the effects of fundamental frequency are reduced

   b) coarser resolution at high frequencies than at low frequencies

   c) easier access to the excitation characteristics

   d) all three were invented by Swedish basketball players

   ⇒All three methods are approaches to estimating smooth short-term spectra, and the smoothing has the effect of reducing the influence of pitch. Thus (a) is the answer. (b) is true for mel cepstra and PLP cepstra, but not for LPC cepstra - the error criterion is equally significant at high frequencies as at low frequencies. (c) is not right either - LPC analysis can directly provide an estimate of the excitation through the LPC residual, but PLP and mel cepstral analysis generate cepstra that result after spectral warping, so the excitation is not easily obtained. It has been pointed out to us that it is more likely that the basketball player inventors were American than Swedish.

6. (2) The frequency response in the peripheral auditory system is tonotopic (meaning center frequency changes with place) because:

   a) the frequency response of the different hair cells varies.

   b) basilar membrane stiffness decreases with distance from the base.

   c) synchrony of auditory fibers decreases with frequency.

   d) none of the above; auditory filter bandwidths are a function of cortical neurons.

   ⇒ b) is correct. For a long time, investigators were baffled by the apparent sharper tuning of neural fibers compared to von Bekesy's original basilar membrane measurements, but recent BM measurements using Mossbauer techniques indicated that BM vibrations are responsible for the tuning.

7. (2) 2400bps was chosen as a standard for secure speech (in the 1950's) because

   a) 2400bps resulted in very good synthetic speech.

   b) it matched the rates required by a teletype when sending real-time transcribed speech

   c) speech was marginal but telephone line modems could not accommodate higher bit rates.

   d) 2400bps could easily be sent down ordinary telephone lines without the need for modems.

   ⇒ c) is correct. In the 1950's, existing phone lines and modems could barely handle 2400bps–so, although the vocoded speech was definitely not great, it was the best compromise.

8. (4) Explain the primary differences between the multipulse, CELP, and VSELP vocoders.

   ⇒ Multipulse, CELP, and VSELP vocoders differ in the way they code the LPC excitation or error signal. In multipulse LPC, the excitation signal is approximated by a set of discrete impulses (typically 4-8) per short segment of excitation (e.g., every 5 ms). An analysis-by-synthesis technique is used to determine the best locations and amplitudes of the impulses.

   In Code-Excited Linear Prediction (CELP), vector quantization is used to code the excitation signal. The codebook consists of random sequences as well as some deterministic sequences in some cases. The coder then picks the codebook vector that minimizes a perceptually weighted error signal when the excitation signal is passed through the LP synthesis filter. The index into the codebook is all that is transmitted for the excitation.

   Finally, Vector Sum Excited Linear Prediction (VSELP) codes the sum of of vectors from multiple (smaller) codebooks to generate the excitation seqeunce. This allows for mixed excitation as well as faster codebook searching. Typically two standard, optimized codebooks and one adaptive codebook are used.

9. (5) The forward recursion to compute model likelihoods can be expressed as

$$\alpha_n(\ell) \;=\; \sum_{k=1}^{L} \alpha_{n-1}(k) P(q_\ell^n, x_n | q_k^{n-1}, X_1^{n-1}, M) \tag{1}$$

where X_1^{n-1} means the sequence of acoustic vectors $x_1, x_2, ...x_{n-1}$, the notation q_ℓ^n means the state at time n with category ℓ, and where there are L different state categories.

As it is commonly implemented, however, it is often expressed as

$$\alpha_n(\ell) \;=\; \sum_{k=1}^{L} \alpha_{n-1}(k) P(q_\ell^n | q_k^{n-1}, M) P(x_n | q_\ell^n, M) \tag{2}$$

Show the steps necessary to go from the first formulation to the second. For each step say whether any assumptions are required for the equality, and if so, say what they are.

$\Rightarrow$ **We begin by factoring the last term of (1) using the definition of conditional probability (no assumptions required here):**

$$\alpha_n(\ell) \;=\; \sum_{k=1}^{L} \alpha_{n-1}(k) P(q_\ell^n | q_k^{n-1}, X_1^{n-1}, M) P(x_n | q_\ell^n, q_k^{n-1}, X_1^{n-1}, M) \tag{3}$$

For the first factor we assume that once the value of q_k^{n-1} is known, knowledge of the previous observations adds no new information. In other words, the current state is conditionally independent of the past observations (conditioned on the previous state). Thus, $P(q_\ell^n | q_k^{n-1}, X_1^{n-1}, M) \approx P(q_\ell^n | q_k^{n-1}, M)$.

For the second factor we assume that once the value of q_ℓ^n is known, knowledge of the previous states and observations adds no new information. In other words, the current observation is conditionally independent of the past states and observations (conditioned on the current state). Thus, $P(x_n | q_\ell^n, q_k^{n-1}, X_1^{n-1}, M) \approx P(x_n | q_\ell^n, M)$.

Given these two assumptions, and substituting the reduced forms into (3), we arrive at the formulation given by (2).

10. (3) Define the following terms (1 sentence per definition):

   a) phoneme

   $\Rightarrow$ **The "alphabet" of acoustic sounds in any language. According to Pickett—"the speech sounds that differentiate words". The phoneme dictionary may differ from language to language but for any one language, all sounds are represented.**

   b) EM $\Rightarrow$ **Expectation Maximization (sometimes called estimate and maximize) - an iterative approach to statistical density estimation for models with missing information - parameters are adjusted to maximize the expected data likelihood over the joint density of known and unknown components.**

   c) analysis by synthesis

   $\Rightarrow$ **A method (often used in speech processing) where synthesis as well as analysis is performed at the analyzer terminal, so that the system can compare the synthesized result with the original speech and adjust the analysis to give the "best"result.**

11. (4) The phrase "carrier nature of speech" was proposed by Dudley as a way of explaining how a vocoder could represent speech using fewer bits (or less bandwidth). Explain how two of the three major algorithms (channel vocoders, LPC vocoders and cepstral vocoders) implement this concept and, as a result, represent the speech signal with fewer bits (or less bandwidth) than a standard telephone channel or a standard PCM system.

    ⇒ In a channel vocoder, the spectrum is represented by the magnitude signals from a bank of bandpass filters covering the speech spectrum of interest. Since this representation varies slowly with frequency, it can be defined by as few as 15 spectral samples. Also, the magnitude signals are slowly-varying in time, so each has a bandwidth of less than 50Hz; thus the spectrum can be represented with about 750Hz.

    In LPC, the synthesizer can be modelled as an all-pole digital filter of approximately 10th order. Empirical results yield a bit rate comparable to that of a channel vocoder.

    In cepstral vocoding, the slowly varying short-time spectral envelope is transformed into the low-time component of the cepstrum, which, again, can be efficiently coded.

    In all three cases, the excitation parameters and the vocal tract parameters are coded separately. For the low rate versions, slowly varying pitch and voicing parameters are detected and coded.

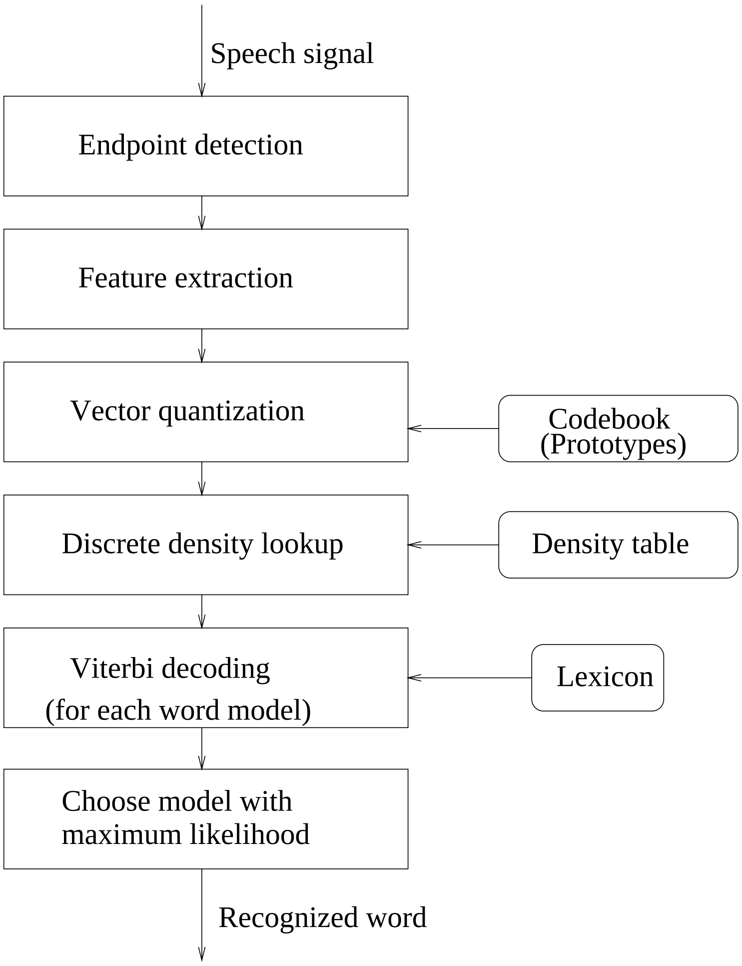
12. (2) Describe at least 3 important steps in a text-to-speech synthesizer

    ⇒ 1. Grapheme to phoneme translation—in English, the sequence of printed letters (or numbers) is transformed into a phonemic description of the utterance.

    2. The derived phoneme sequence is transformed into a varying set of parameters to control a specific speech synthesizer.

    3. Synthesizer "speaks" in response to the applied temporal fluctuations of the parameters.

13. (4) Draw a block diagram of an isolated word speech recognition system based on discrete probability distribution Hidden Markov Models. Explain the function of each of the constituent parts.



⇒ The speech endpoints are found (to reduce false matches to between word models and non-speech sounds), and then for each frame step (typically every 10 ms) a feature vector such as PLP cepstra is computed. It is quantized to the nearest vector prototype from the codebook (multiple codebooks are often used for multiple feature types, such as cepstra and delta cepstra). The prototype index is used to look up a discrete density function, i.e., the probability of any state given the vector prototype. This density is used in the decoding process for each word (model) in

the lexicon; typically a Viterbi criterion is used for the decoding. The word corresponding to the model with the maximum (Viterbi) likelihood estimate is chosen. Not shown but possible (even for isolated words) is an additional multiplicative factor for the prior probability of each word.

14. (4) Describe at least two distinct ways in which vector quantization (VQ) is used in different types of vocoders. Briefly compare this to their use in discrete density HMM-based speech recognizers. (I.e. are there similar reasons that VQ is effective in both cases? Is it useful for similar reasons?)

    $\Rightarrow$ **In several types of vocoder (channel, formant, LPC, etc.) the spectral envelope information in the form of a feature vector (e.g., formant or cepstral parameters) is mapped to the nearest vector in a codebook of prototype entries during analysis. An index to this "best" vector is sent through the communications channel, and during synthesis the index is used to look up the parameters that will be used to reconstitute the speech. Examples include the channel vocoder from C.P Smith's experiments, in which vector patterns of filter magnitude functions were stored, with the intent of including all patterns leading to perceptually distinct speech. Another historic example was the Kang-Coulter LPC-based 600bps formant vocoder, in which the different patterns of three formants were stored as vectors. This required a much smaller storage capacity than for the channel vocoder case due to the formant representation.**

    In other algorithms, most notably CELP and VSELP, VQ is used to quantize the excitation signal - in fact, "Code" from the CELP acronym refers to a VQ codebook of excitation prototypes.

    For discrete density HMM-based speech recognition, the quantization is similar to that done in the first application above - to succinctly represent the spectral envelope (short-term spectrum) or related quantities such as the short-term cepstrum. In all of these cases, VQ can provide such a succinct representation with little performance reduction (intelligibility or recognition performance, depending on the application) because there are a range of such vectors that are functionally equivalent for vocoding or recognition, and the range of probable vectors is much smaller than the range of all possible vectors.