EECS 225d: Midterm 2, April 24, 1995

Version with correct answers (not necessarily unique)

Name:

For multiple choice and true/false 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) The Baum-Welch procedure is

   a) a statistical algorithm for recognizing with an HMM

   b) a statistical algorithm for training an HMM **X**

   c) an approach to filter design

   d) a kind of cosmetic surgery

   The Baum-Welch procedure is a training algorithm which is an application of EM to HMMs (choice b). Parameters are iteratively modified to improve a likelihood criterion, using probability estimates derived from the previous set of parameters. Forward and backward recurrences are used to estimate the required probabilities.

2. (2) The pitch of a harmonic series with missing fundamental frequency is close to

   a) the fundamental frequency **X**

   b) twice the fundamental frequency

   c) halfway between a) and b)

   d) none of the above

   e) a fastball

   The pitch of a harmonic series with missing fundamental is close to a) the fundamental frequency. This has been established through numerous psychoacoustic tests.

3. (10) Briefly describe the different approaches to speech synthesis by 5 of the the following:

   a) Dudley and his Voder

   b) John Holmes and his work with Fant's OVE II Synthesizer

   c) The Haskins Laboratories researchers using the pattern playback.

   d) George Rosen and his Vocal Tract Synthesizer.

   e) Dennis Klatt and his DecTalk.

   f) Synthesis by rule

   g) Sid Fels and GloveTalk II

   Include such topics as motivation, implementation, synthesis strategy.

   a) Dudley's important concept was "the carrier nature of speech". In later years "convolution" was used to describe the same idea; namely, that speech could be described as the convolution of a wide band excitation signal with a band restrictive, time varying linear filter. He used these ideas in the design of the Voder and Vocoder. In addition, the Voder activity included the very important intensive training of the performer to learn to manipulate the keyboard properly.

   b) John Holmes first carefully analyzed an English sentence in terms of the parameters of the OVEII speech synthesizer of Gunnar Fant's laboratory in Stockholm. He then manipulated the parameters of OVEII and by trial and error eventually generated a very good replica of the analyzed speech. He thus proved that under certain conditions this synthesizer could synthesize essentially perfect speech. By tabulating the information content of the parameters he used, he was able to demonstrate the "ideal" effectiveness of this system as the receiver in a bandwidth compression scheme.

   c) The pattern playback was an optical device that tranformed a hand-made spectrogram into speech. Pulses of light were obtained from a rotating disc and a photocell transformed the reflected light into sound. The main focus of the research was to examine the cues for speech perception, (primarily of the plosive consonants) in terms of the hand-drawn formant transitions. Despite the poor quality of the speech from this rudimentary synthesizer, many valuable results on speech perception were obtained.

   d) Rosen's Vocal Tract Synthesizer was an early attempt, using analog hardware, to physically model the acoustic properties of the vocal tract.

e) Klatt's Text-To-Speech synthesizer was the culmination of many years of research that included a high quality synthesizer but many other basic items such as letter to phoneme rules (this also dealt with the many exceptions in English),prosody rules and interface protocols. This device, in software, was of great value to research workers who needed to generate synthetic speech for their work. (An interesting and notable example is the Miller and Sachs study of the cat's auditory system).

f) Synthesis by Rule is a method of deriving synthetic speech by specifying as input a phoneme string.

g) GloveTalkII is a system for producing speech using hand motions and a specially designed glove. It is perhaps the closest thing to von Kempelens synthesizer in that it utilizes hand gestures-to-parameter transformations. It is related to the Voder in the sense that much training is required to make intelligible sounds.

4. (4) Auditory experiments have determined that an LPC vocoder of order eight is about as good as a channel vocoder that requires 14-16 channels. Briefly present an argument that explains this result.

   LPC analysis results in an all-pole synthesizer and this can be interpreted as a model with a reasonably precise pattern of formant frequencies. An eight channel vocoder, on the other hand, would need wide filters to cover the band. A channel vocoder synthesizer is a pole-zero system and the poles are fixed (and non-variable) by the filter design, so it can adapt to the spectrum only by varying the zeros and this, in turn is done by controlling the modulating signals derived from the analysis filter bank. By the way, this does not imply that LPC bit rate can be less than channel vocoder bit rate for equal quality. In addition to the NUMBER of parameters, total bit rate also depends on the number of required QUANTIZATION levels needed for effective transmission.

5. (2) It has often been stated that female speech is more difficult to vocode than male speech. Do you believe this? Explain.

   For channel vocoders and LPC vocoders this seems to be the case. A plausible way to look at this issue is this: the speech spectrum can be thought of as an envelope sampled at the harmonics of the fundamental frequency (at least this model works for the voiced sounds). Since, on the average, female pitch is higher than male pitch, sampling is sparser so that spectral reconstruction for synthesis is poorer.

6. (4) User A has an available channel vocoder and user B has an LPC vocoder. Can they communicate without additional hardware? If so, explain; if not, sketch what extra hardware would be needed.

   If user A talks through a Vocoder analyzer, the parameters received by user B cannot be used by his LPC synthesizer without further processing. One way of making the link A-B compatible would be for User B to compute the square of the received spectrum, then compute the correlation function by taking the DFT of that square and finally doing an LPC analysis from this derived correlation function. User A could build an LPC synthesizer, pulse it to get the impulse response and take a DFT of that to get the spectrum envelope. Although this seems stupid since, if A has the LPC synthesizer, why not use it directly to generate the synthetic speech? But A may feel (rightly or wrongly)that his channel vocoder synthesizer is a higher quality synthesizer even if it gets its information from an LPC analysis.

7. (2) If random variable q corresponds to class and x is the feature vector, the minimum probability of error will be achieved by always choosing the value for q that maximizes

   a) $p(q)$

   b) $p(q)p(x|q)$ **X**

   c) $p(q)p(q|x)$

   d) $p(x)$

   The minimum error decision is to choose the class with the largest $p(q|x)$. Using Bayes' rule, this is equivalent to to choosing the class with the largest value for $\frac{p(q)p(x|q)}{p(x)}$. Since $p(x)$ is the same for every class, the denominator can be ignored, leading to b).

8. (2) Give a statistical condition that would make a linear discriminant function the optimal (minimum probability of error) classifier for a 2-class problem.

   If the densities are Gaussian with equal covariance matrices for every class, the optimal classifier is a linear discriminant. See Duda and Hart.

9. (2) For a minimum distance classifier, give a reason why Euclidean distance might not be the best one for classification.

   If feature vector components are different in size (e.g., one has much larger variance than the others), then they will have an unbalanced amount of influence on the total distance, which may not correspond to their importance or reliability for discrimination. Even if the variances are equal, components may differ in terms of their relevance for discrimination.

10. (4) Give 2 methods of time normalization between reference word templates and input word templates, giving at least one advantage of each.

    1. DTW: dynamic time warping, which provides an optimal path through the reference templates (in the sense of the particular distance metric chosen), permitting locally variable compression or expansion of time.

    2. Linear time compression/expansion: this has the advantage of simplicity in comparison with DTW, but obviously cannot account for anything but a single average stretching or compression over the whole utterance.

    Some people also referred to HMMs, which are not usually used with templates per se, but they also provide time warping.

11. (2) What is a Markov model?

    A Markov model is a probabilistic finite state machine (that is, a finite state machine with probabilistic transitions), such that the probability of moving to a state is only dependent on the previous k states (for a $k-th$ order Markov model). The most common Markov asssumption is the 1st order one, so that the probability of the move to a new state is only dependent on what the previous state was.

12. (6) Draw a block diagram showing the major components of an isolated word speech recognition system, and explain some reasonable design choices for each major piece. Include an explanation of how the system would be trained.

    Endpoint detection is typically required for isolated word recognition. This can be done in a number of ways, but usually one must set thresholds in both amplitude thresholds and time.

    Once the start of word is detected, feature vectors are calculated every 10 msec (a common stepsize). In practice, of course, the features may be calculated continuously and used for the endpoint detection. A reasonable front end would consist of PLP cepstral coefficients and their corresponding delta (1st derivative) parameters.

    Pattern classification can be done using dynamic time warp between prototype reference templates and incoming templates formed by the sequence of vectors between templates, using a Mahalanobis distance. Training can be done by clustering many occurences of the references into a single prototype per vocabulary word. DTW can be used for the clustering as well, since the examples will typically be of different lengths. The covariance matrices for the features can be computed from the full training set in order to evaluate the distances.

    Each of the choices above was, of course, one of many possible.

Answer 3 of the following 5 remaining questions.

13. (6) At different times, the dominant frequency region for pitch perception has been considered to be the region of the fundamental frequency, the region of the 10th harmonic, and the region near harmonics 3, 4, and 5. Discuss the history of these hypotheses in connection with the ideas and experiments of Helmholtz, Schouten, Licklider, and Ritsma.

    Helmholtz theorized that a person's perception of pitch is a function of the PLACE on the basilar membrane that is stimulated by the tone.This implied that the fundamental frequency of a complex tone had to be present in order to perceive its pitch. When his contemporary Seebeck showed that the fundamental frequency was perceived to be the pitch even if it were absent, Helholtz responded by proposing that a non-linearity in the ear caused generation of this missing fundamental. Many years later, Schouten proposed a temporal theory of pitch perception wherein neural volleys occurring at relatively high frequencies were perceived as pitch by the auditory system. Licklider augmented this theory by performing a psychoacoustic experiment in which he used a noise masker to mask out any possible auditory response at the fundamental frequency. Since the fundamental frequency could still be perceived although not present, this showed that the Helholtz conjecture was mistaken. Ritsma searched for that part of the spectrum where pitch perception would dominate. His work resulted in the realization that perception of pitch, unlike Schouten's conjecture, depended on the low order harmonics of a complex tone.

14. (6) Answer the following about Fletcher's Articulation experiments as presented by Jont Allen:

    a) Why did Fletcher use nonsense CVC's as the database?

    b) What was Fletcher's rationale for defining Articulation bands?

    c) What is the difference between Articulation and Recognition?

    a) Fletcher did not want his results to be influenced by context.

    b) On the average, the error obtained for equal articulation bands were equal—but note-the bandwidth of equal articulation bands was not constant but increased with frequency.

    c) Articulation is recognition of NONSENSE CVC's–Recognition is for real words.

15. (6) Why is spectral flattening of a speech signal followed by autocorrelation a good front end for pitch detection? Define autocorrelation and spectral flattening in your answer.

    To a first approximation, spectral flattening yields equal amplitude harmonics. Correlation produces equal phase among all harmonics.The resulting spectrum resembles that of a periodic pulse train—for such a signal the period should be easy to ascertain.

16. (6) Discuss the concepts behind the Pythagorean, Just, and Equal Tempered Scales. Which of these scales is prevalent in western classical music? Why? What is the Pythagorean comma and why is it a problem?

    The Pythagorean scale is obtained by using the Circle of Fifths to generate all the notes of the scale, and assigning a frequency for the new note equal to 3/2 of the previous note. When all 12 notes have been assigned a frequency, the next note, which is C, 7 octaves above the starting note will have a frequency of (3/2) raised to the 12th power, which is SLIGHTLY greater than 2, raised to the 7th power. This increment is the Pythagorean comma. The just scale is obtained by assigning the ratios 4:5:6 to the major chords. The equal tempered scale is based on the idea of equal ratios for successive notes of the scale. Thus, for example, in the 12 tone chromatic scale the frequency ratio of C# to C is the same as E to F, G to G#,etc. This scale has the advantage that all melodic intervals have the same ratio in all keys. A possible disadvantage is slight deviation of equal tempered notes from the just scale and MAY be perceived by very sensitive ears as slightly dissonant.

17. (6) A large number of spoken sentences are recorded, stored, shuffled into random order, and split into two groups. The waveforms for one of the groups is processed with a linear time invariant filter. If the unfiltered group is used for training a speech recognizer and the filtered group is used for testing it, how can you choose feature vectors that reduce the effect of this mismatch, without explicitly learning this filter characteristic and compensating for it? Describe any of the techniques we discussed (or any other you can think of that would help).

    The easy answer for this was either cepstral mean subtraction (CMS), RASTA, or delta cepstra. The main point is that the effect of the LTI filter is to convolve the speech with an impulse response, which corresponds to multiplying the power spectrum by some function, or adding another function to the log power spectrum (or, equivalently, to the cepstrum). This added function is constant over the speech, while the speech component is time-varying, so that a mean subtraction, derivative, or highpass or bandpass filtering of log spectral or cepstral trajectories gets rid of (to 1st order) this time-invariant additive component.