音響音声学・冬学期レポート課題

〆切：2 月 3 日（日）23 時 59 分 59 秒　提出先：mine@gavo.t.u-tokyoac.jp までメールで

第 1 問　音声認識では，スペクトル包絡ではなく，ケプストラム係数を使うことが多い。まず，ケプストラム係数とは何かを述べ，次に，何故（スペクトル包絡ではなく）ケプストラムが用いられるのかについて述べよ。

第 2 問　Dynamic Time Warping とはどのような技術なのか，知るところを述べよ。

第 3 問　隠れマルコフモデル（Hidden Markov Model）とはどのような技術なのか，知るところを述べよ。

第 4 問　HMM は音声認識の音響モデルとしても，音声合成の音響モデルとしても使われている。基本となる技術は同じであるが，その使い方，使う条件は様々な違いがある。どのような違いがあるのか，知るところを述べよ。

第 5 問　音声認識は，入力音声とシステムが有する複数のテンプレートとの音響的比較（音響的照合）が基本となる。その昔，音響的照合は DTW が主流であった。その後，音響的照合は，入力音声（ケプストラム系列）と HMM を比較することが常識となった。前者に対して後者はどのような強みを持つのか？説明せよ。

第 6 問　音声認識は，HMM などによる音響モデル以外にも，どのような単語が次に来るのかを予め予測する言語モデルの助けが必要である。言語モデルの構築の仕方（次単語の予測の仕方，次に来る単語の絞り込み方）として大きく二つの方法を説明した。各々について説明せよ。

第 7 問　ベイズの定理を使って，条件付き確率を下記のように変形することが技術的に広く行なわれている。まず第一式の $\square$ を埋めよ。第二，第三の $\square$ の間に縦棒があることに注意せよ。次に，第二式が成立することを第一式を使って説明せよ。最後に，第二式の意味するところを，音声認識の文脈で説明せよ。即ち o = 観測された（入力される）ケプストラム系列，w = 単語列，として説明せよ。

$$P(w|o) = \frac{P(o|w)P(w)}{P(o)} = \frac{P(o|w)P(w)}{\sum_{w'} P(o,\square)} = \frac{P(o|w)P(w)}{\sum_{w'} P(\square \mid \square)P(\square)}$$

$$\underset{w}{\operatorname{argmax}}\, P(w|o) = \underset{w}{\operatorname{argmax}}\, P(o|w)P(w)$$

第 8 問　更に s = 話者，とした場合，次の二式の意味するところを説明せよ。

$$P(o|w) = \sum_s P(o,s|w) = \sum_s P(o|s,w)P(s|w) \approx \sum_s P(o|s,w)P(s)$$

$$P(o|s) = \sum_w P(o,w|s) = \sum_w P(o|s,w)P(w|s) \approx \sum_w P(o|s,w)P(w)$$

第 9 問　HMM 音声合成の場合，文 HMM（長〜い HMM）から尤度が最大となるようなケプストラム系列を得ることができるが，ナイーブに実装すると，状態 i の（分布 i）の平均ケプストラムが数フレーム出力され，次に，状態 $i+1$ の平均ベクトルが数フレーム出力され，と階段状のケプストラム時系列が生成されることになる。これでは自然な合成音声は得られない。スムーズなスペクトルパターン（ケプストラム時系列パターン）を得るための工夫について知るところを述べよ。（必ずしも数式を使うことは要求していない）

第 10 問　最後の数回は，音声の構造的表象について，a)（発達，進化という軸を見据えた）その理論的背景，b) 数学的定式化，c) 実用アプリ・システムへの応用例，d) そして言語の起源に関する妄想について述べた。以下の三つの作文（web から入手できる）を読み，音声の構造的表象について，自由にコメントを述べよ。各々異なることが述べられているので，コメントを読めば，どれを読んでないかは一目瞭然となることに注意せよ。

「音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案」
「声とは言葉とは何か 〜音声研究を通して考えること〜」
“A MODULATION-DEMODULATION MODEL FOR SPEECH COMMUNICATION AND ITS EMERGENCE”