Department of Electronics and Telecommunications

Examination paper for TTT4185 Speech Technology

Academic contact during examination:: Tor André Myrvoll
Phone: +47 95 14 80 14

Examination date: Thursday 3. December 2015

Examination time (from - to): 09.00 - 13.00

Permitted examination support material: **C** – Specified, written and handwritten examination support materials are permitted. A specified, simple calculator is permitted

Other information:

- The examination consists of 3 problems where
  - problem 1 concerns automatic speech recognition
  - problem 2 concerns speech analysis
  - problem 3 concerns speech synthesis
- All sub-problems counts the same
- All problems are to be answered
- Grades will be announced 3 weeks after the examination date.

Language: English

Total number of pages: 10

Of this, number of enclosure pages:

Checked by:

Date Signature

Note! Students will find the examination results in Studentweb. If you have questions about your results, you must contact your department. The Examination office will not be able to answer such inquiries.

Problem 1

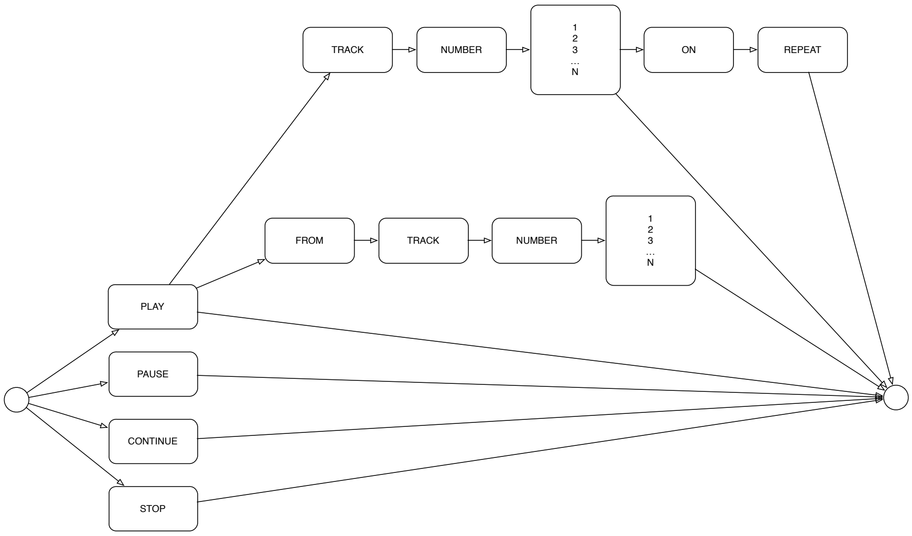
You have been assigned the task of designing a spoken language interface for controlling the CD-player in a new car model. The prototype must recognize the following command phrases

- Play/Pause/Continue/Stop
- Play track number N
- Play track number N on repeat
- Play from track number N

where N is a number between 1 and 30.

1a) Design a finite state machine (FSM) grammar that represents *only* the allowed command phrases. The FSM will start in the state START and end in the state FINISHED. One can assume that all the numbers are found in the utterance lexicon, so designing a grammar for these is not necessary. The state of the CD-player is also of no concern (It is ok to recognize the phrase 'Pause' even when the player is not running and so on).

A suggested solution is found in the figure below.



1b) Speech recognizers can be based on different speech units like words, syllables, phonemes and context dependent phonemes. Common to all are the concepts of accuracy, trainability and generalizability. Explain these concepts and how they apply to words, phonemes and context dependent phonemes.

There are mainly three characteristics we want a speech unit to have: accuracy, trainability and generalizability.

- Accuracy: The unit should not be easily confused with other units.
- Trainability: There should be enough data to train the unit models.
- Generalizability: New words should easily be infered from the units.

- Words
  - High accuracy, since the all coarticulation effects are internal to the word.
  - Low trainability since the number of words are is very high.

  - Low generalizability as a new data needs to be collected and models trained if new words are to added
  - Suited for tasks with small vocabularies
- Phones are the most common speech unit in current systems, and can be further divided into two types:
  - Context independent phones
    * Low accuracy since coarticulation effects are not modeled.
    * Very high trainability since the number of phonemes is low (40-50).
    * Very high generalizability.
    * Suited for tasks with little available training data and little information about the kinds of input the system will receive.
  - Context dependent phones
    * High accuracy since coarticulation effects over triphones are modelled.
    * Low trainability since the number is very high. (However, clustering states where different contexts give rise to similar coarticulation effects can reduce this number considerably.)
    * Good generalizability.
    * Well suited for all tasks where there is sufficient data available to train them.

1c) Which speech unit is most suitable for the CD-player user interface? Discuss the pros and cons of your choice.

1d) Explain the concept of hidden Markov models (HMM), and how they can be used to model speech units.

1e) A database of spoken utterances covering the chosen speech units has been recorded, and a small subset has been annotated with the start and stop times of the speech units. Give a high level description of the procedure of training acoustic models using these data.

- Use the transcribed data to create single mixture monophone models.
  - The start, middle and end states are initialized by dividing every transcribed phoneme into three equal parts and then computing the mean vector and covariance matrix.
  - Use forward/backbard algorithm to update the monophones until convergence.
- Convert sentences to phoneme strings using the pronunciation lexicon.
- Use forward/backbard algorithm to update the monophones until convergence.
- Increase the number of mixtures in the monophone models
  - Split every mean vector in two, effectively double the number of mixtures or,
  - Split the mean vector of the mixture with largest mixture weight, adding one mixture
  - Use forward/backbard algorithm to update the monophones until convergence.

It has recently been brought to the attention of the company that absolutely nobody buys CDs anymore, and that the future (for now) belongs to large libraries of audio files. A more ambitious speech interface that recognizes natural speech is now being considered. Command phases may include

- Play the album Washing Machine by Sonic Youth
- I would like to listen to some Jazz

1f) FSM grammars are no longer a viable alternative. Instead, a large vocabulary recognizer with a statistical language model will be utilized. Describe the general n-gram language model and express the probability P('I would like to listen to some Jazz') using a 3-gram (trigram). The n-gram model models the probability of a given word in an utterance as a discrete probability given the $n-1$ previous words in the utterance.

$$\begin{aligned}&P(\text{I would like to listen to some Jazz})\\ =&P(\text{I})P(\text{would}|\text{I})P(\text{like}|\text{would, I})P(\text{to}|\text{like, would})P(\text{listen}|\text{to, like})P(\text{to}|\text{listen, to})\\ &\times P(\text{some}|\text{to, listen})P(\text{Jazz}|\text{some, to})\end{aligned}$$

1g) It is assumed that a vocabulary of 40000 words will cover most of what is needed for the speech interface. How many parameters does one in principle have to estimate for 1-, 2- and 3-gram models? Explain the motivation for using *discounting*, *backoff* and *interpolation* when training n-gram language models, and describe the three techniques briefly.

- 1-gram: 40000 parameters
- 2-gram: $40000^2 = 1.6 \times 10^9$ parameters
- 3-gram: $40000^3 = 6.4 \times 10^{13}$ parameters

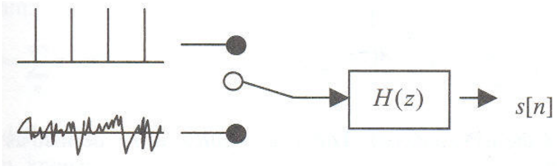
Discounting, *backoff* and *interpolation* are used to address the problem of estimating eg. 64 trillion parameters for a trigram (there is just not enough data in the world. Also, you would want to store all those parameters). In practice most trigrams would have zero examples in a limited training set, giving them a probability of zero, which would be a problem if that trigram was encountered during use.

- Discounting: Keep a small probability mass in reserve and distribute it evenly across all unseen trigrams.
- Backoff: If the trigram isn't found – use the bigram. If the bigram is not found – use the unigram.
- Interpolation: $P_I(w_3|w_2, w_1) = \alpha P(w_3|w_2, w_1) + \beta P(w_3|w_2) + (1 - \alpha - \beta)P(w_3)$ The weights $alpha, \beta$ are usually trained to maximize perplexity.

Problem 2

2a) Sketch a simple source-filter model for speech and describe the constituents.

The source-filter model:



Time varying filter represents transfer function of vocal tract. Excitation is pulse train (voiced sounds) or pseudo-random noise (unvoiced sounds). The gain control models volume (or loudness).

Physically, the voiced sounds are produced by periodically opening the vocal cords to let a pulse of air through, while unvoiced sounds are produced by constrictions in the vocal tract (or at the lips), resulting in a turbulent, noise-like air-flow.

It is well suited for Voiced and unvoiced phonemes, but not so much for plosives.

2b) A common problem in speech analysis is to separate the source and the filter given a set of samples from a speech signal. Explain how linear prediction can be used to model the filter (no detailed mathematical derivation of the estimator is expected). Also explain how the source signal can be estimated given a set of filter coefficients.

A linear predictor of order p is given as

$$\hat{y}(n) = \sum_{k=1}^{p} a_k y(n-k)$$

and is defined by the LP-coefficients $\{a_k\}$. We can also define the prediction error

$$e(n) = y(n) - \hat{y}(n)$$

which in turn means that we can write $y(n)$ as

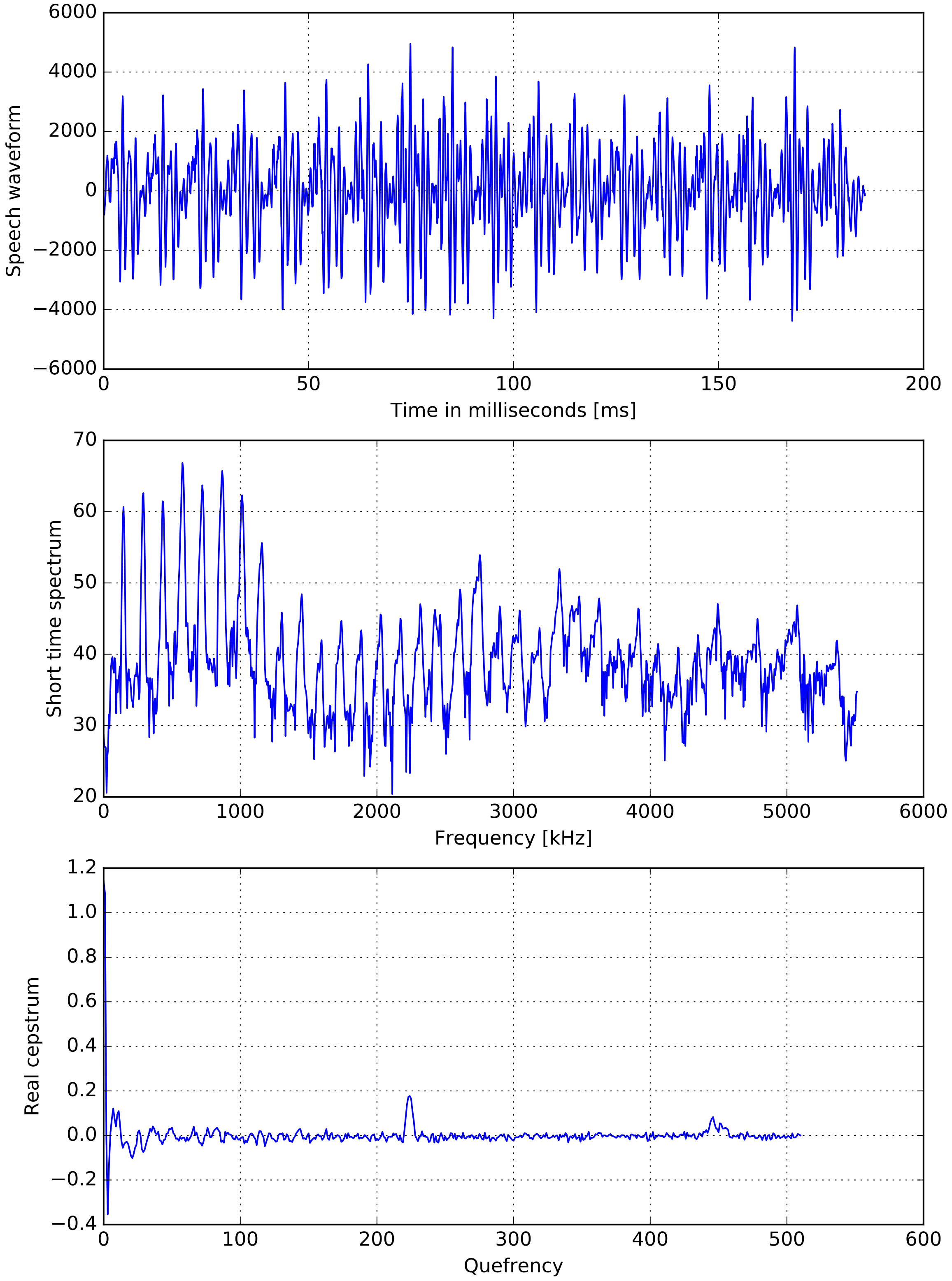
$$y(n) = \hat{y}(n) + e(n) = \sum_{k=1}^{p} a_k y(n-k) + e(n)$$

which is a filter of order p.

To obtain the LP-coefficients we minimize $E[e^2(n)] = E[(y(n) - \hat{y}(n))^2]$, where $E[\cdots]$ is the expectation, which usually is approximated by the sample mean. Differentiating wrt. $\{a_k\}$ and solving the linear equation yields the LP-coefficients.

To estimate the source signal we simply compute $e(n) = y(n) - \hat{y}(n)$ given the estimated LP-coefficients.

2c) Explain what a *formant* is and describe the two parameters describing it. What order should a linear prediction filter have to approximate spectrum containing three formants?


6000
4000
2000
0
−2000
−4000
−6000
Speech waveform
0
50
100
150
200
Time in milliseconds [ms]
70
60
50
40
30
20
Short time spectrum
0
1000
2000
3000
4000
5000
6000
Frequency [kHz]
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Real cepstrum
0
100
200
300
400
500
600
Quefrency

Formants are the acoustic resonances of the speaking apparatus, or the spectral peaks of the sound spectrum of the voice. They are specified by their position (of the the peak), and their width.

Minimally, one needs to specify the position, in which case a 6-order filter is sufficient (2 poles per formant). If the width is important, a minimum of 12 coefficents are needed.

2d) What is the motivation behind short-time Fourier analysis applied to speech signals? Which tradeoffs are typically made when choosing window functions? Describe the pros and cons of the rectangular window function.

Short-time Fourier transforms are useful for speech processing because the speech signal is stationary over short timescales, typically 25 milliseconds.

Choosing a window-function is in general a tradeoff between resolution (determined by the width of the main lobe of the window in the frequency domain), and smearing (determined by the attenuation of the side-lobes).

The rectangular window has the most narrow main-lobe, but the poor attenuation of the side-lobes introduces a lot of smearing.

2e) The real cepstrum is defined as

$$c(k) = IDFT\{\log|X(\omega)|\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)|e^{j\omega k}d\omega$$

where $X(\omega)$ is the Fourier transform of the signal $x(n)$. Explain how the cepstrum kan be used to separate source and filter for a speech signal.

In the frequency domain the speech signal is a product of the source and filter. Using the log-function we can show that

$$\begin{aligned} c(k) =& IDFT\{\log|Y(\omega)|\} && (1)\\ =& IDFT\{\log|X(\omega)H(\omega)|\} && (2)\\ =& IDFT\{\log|X(\omega)| + \log|H(\omega)|\} && (3)\\ =& IDFT\{\log|X(\omega)|\}IDFT\{\log|H(\omega)|\} && (4)\\ =& c_X(k) + c_H(k) && (5) \end{aligned}$$

Since the filter and source are slowly and rapidly changing respectively, it is reasonable to assume that there exists some K where $c(k)$ for $k < K$ describes the filter, and vice versa for the source. Eg. setting $c(k) = 0\ \forall k > K$ and then transforming the cepstrum back to the spectral domain will give an estimate of the filter.

2f) The figure on the next page shows three different speech signals in the time, frequency and cepstral domains respectively. All of the signals were sampled at 44100 Hz. Estimate the pitch for the three signals, and explain the approach used. Note: The pitch may *not* be the same for the three signal.

- Waveform: We see we have approximately 10 pulses per 100 milliseconds, which in turn yields $T_0 = 10$ milliseconds and $F_0 = 100$ Hz.
- Spectrum: There are about 7 peaks corresponding to the source in the first 1 kHz band. We get $F_0 = 1000/7 = 142$ Hz.

- Cepstrum: There is a clear peak in the cepstrum at about index 220. This means there is a strong component in the spectrum that cycles 225 times in 44100 Hz. This yields $F_0 = 44100/225 = 196$ Hz.

Problem 3

3a) Which four main processing blocks constitutes a text-to-speech system? Briefly describe the function of each of the blocks.

- Text analysis: text normalization; analysis of document structure, linguistic analysis

  Output: tagged text
- Phonemic analysis : homograph disambiguation, morphological analysis, letter-to-sound mapping

  Output: tagged phone sequence
- Prosodic analysis: intonation; duration; volume

  Output: control sequence, tagged phones
- Speech synthesis: voice rendering

  Output: synthetic speech

3b) *Large Number of Rare Events* is a common issue in concatenation synthesis. Briefly explain the nature of the problem. Is it also a problem for diphone synthesis?

- Large number of units with small probability of occurrence
- If database units are selected randomly, the probability of encountering a unit not in the database approaches certainty for a small sequence of randomly selected sentences.
- Unit inventory must be chosen with care
- Fall-back solutions must exist for non-covered units

For diphone synthesis there is no problem as there are fewer units and prosody are done using signal processing.

3c) In concatenation synthesis speech units are selected from a library of units and concatenated to a speech signal. Briefly explain how the units are chosen given a target sequence. Describe briefly the significance of the transition and unit costs

Given a target sequence $t_1, t_2, \ldots, t_N$, set of units is chosen that minimizes the sum of the

- transition cost $d(u_n, u_n + 1)$, which measures the mismatch between units at the point of concatenation
- unit cost $d(u_n, t_n)$, which measure the mismatch between the target unit and the candidate unit.

3d) Describe the PSOLA algorithm and explain how it is applied to

- Pitch adjustment
- Speaking rate adjustment

PSOLA can modify pitch and duration. The starting point for PSOLA is having accurate estimates of the fundamental frequency in the voiced parts of the speech signal. For every vocal pulse you can then construct a waveform segment centered around the pulse. The segments are attenuated towards the end points and typically extend over two fundamental periods (2T0). Through the additive combination of partially overlapping segments, where the degree of overlap is such that the distance between successive vocal pulses is equivalent to the desired fundamental frequency, a voice signal with the same spectral envelope but a new fundamental frequency is constructed. By repeating or skipping segments from the original signal, the manipulation can be done without changing the duration of the signal.

The duration of the signal can also be changed by keeping the fundamental frequency constant, but instead inserting or deleting segments.