Analytics for Experts
or
For Computer Vision People: How to talk to a Speech Researcher
For Speech People: How to talk to a Computer Vision Researcher

Gerald Friedland
International Computer Science Institute
Berkeley, CA

Version 1.1

Abstract
Even though the mathematical foundations are very similar, in the past, artificial intelligence research seemed to be strictly divided according to the different types of data that were to be analyzed. Therefore many research groups work on either speech processing, computer vision, or video analysis. Only recently trends, such as Semantic Computing, have emerged that try to merge the different research tracks in order to create unified approaches that can benefit from the synergy of extracting and analyzing data of different modalities in a combined way. The following document tries to provide a quick and dirty introduction to the different methodological, philosophical, and foremost terminological branches that have been taken by speech scientist and by computer vision scientists.

Warning:
This document contains strong positions and opinions that are biased by my personal experience. This document is intended as a primer for people wanting to interact with vision or speech researchers. It is not a research paper. As such, this document has no chance of ever being complete. However, I want invite everybody to add to it so it can grow organically.

1. Introduction
The foundations of audio, speech, image, and video analytics are rooted both in the signal processing community, which is part of electrical engineering, as well as the field of statistics, which is part of mathematics. Many terms that are still used have been inherited by these two fields. A newer field, which has come up with the raise of the computer, is called “machine learning”, it is actually a subfield of statistics, paired with biology though. The field of machine learning redefines many words originally created in statistics and biology. The application of machine learning together with the foundations of signal processing to different kinds of data created the different fields of audio, speech, image, and video processing, which by themselves created new terms. There are different reasons for those fields to have become separated, some are social and very pragmatic. A very important one though is the amount of data that has to be processed. Audio and speech processing is a the oldest field because computers where already able to process speech in the 60s and 70s. Image analysis is a slightly newer
field, and video analysis is the newest because there is much more data to be processed. With different maturities of the fields, different generations of people have worked on the different types of data and hence, different vocabulary is used.

Today, the processing power of modern computers is just starting to allows us thinking about approaches that analyze a given video multimodal, i.e. processing the audio, the images, and the dependencies between the images synergistically. A combined processing promises improved robustness in many situations and is closer to what humans are doing. The human brain takes into account not only patterns of illumination on the retina or periods of excitation in the cochlea, it also combines different sensory information and benefits from past experience. Humans are able to use context information and to fill in missing data by associating parts of objects with already learned ones.

This document tries to contribute to the reunification of the research fields by summarizing the most important concepts in each area and the vocabulary used to describe them.

2. Basics of Machine Learning

The basic workflow of a machine learning approach is illustrated as follows:
Using training data, relevant features are extracted and passed to a machine learning algorithms. Examples are: Gaussian Mixture Models, Neural Networks, Support-vector Machines, Decision trees (for example ID3), k-Nearest Neighbors, Bayesian Networks, Hidden Markov Models. Feature extraction methods are usually derived from signal processing or electrical engineering (sensor input/output). Machine learning algorithms, are usually methods derived from mathematical statistics and produce statistical models. These are basically sparse representations of the data that allow thresholding of any kind. In order to train the models, the right answers have to be provided which are given in the ground truth data. This is usually metadata created by human annotators.

In testing mode, the statistical models are then used to perform the actual pattern recognition task. The test data is of course run through the same feature extraction process. The results are either just used or compared against the ground truth in order to benchmark the quality of the algorithm.

Supervised machine learning works just as described. Unsupervised machine learning omits the training step, statistical models are created on the fly using the test data.

3. Speech Processing
With audio being only one-dimensional data, and speech actually being able to be captured at low sampling rates (mostly 8kHz and 16kHz, due to Nyquest theorem) and low bit-resolution (8bit or 16bit, often quantized using a perceptual scale, such as a-law or mu-law) it constitutes the oldest research field of the ones presented here. This has several consequences:

a) Speech processing has the most advanced machine learning techniques.
b) This field can computationally afford the most advanced statistical methods.
c) Scientific progress is generally slower (“in Automatic Speech Recognition, 0.5% improvement is a PhD thesis”).
d) Speech processing has the most advanced benchmarking and testing culture.
e) The majority of the approaches do not seek to work online (i.e. realtime and incrementally as new data comes in).

3.1 Sub fields
The vocabulary in Speech Processing is pretty much dominated by its main subfield: Automatic Speech Recognition or ASR. Speech recognition is sometimes wrongly called “voice recognition”. Only recently, other fields have emerged, such as: Speaker Recognition (given a sample of a speaker is a the given audio recording actually from the same speaker?), Speaker Identification (Given an audio recording and a database os speaker models: “Who is in the recording?”), Speaker Diarization (Given no prior information: “Who (speaker a, b, c,...) spoke when?”), Language and Dialect Identification (Given a database of sample speech in different languages and dialect, the question is “which language is spoken?” in the test sample, Spoken Language Retrieval (Given a database of speech, different questions are targeted: “Where does speaker x say something?”, “When is sentence xyz uttered?”), objective speech and audio recording quality measures (PEAQ, PESQ).
Speech synthesis is a very important field too. However, since mostly no semantic analytics is performed, it won’t be covered by this document (at the moment).

3.2 Non-Speech
Although audio can consists of various things other than speech, such as music and noise, speech processing is still the most dominant field. The discrimination of speech and non-speech (speech activity detection) is a research field on its own. Noise classification deals with the classification of different non-speech signals.

Music processing is a an entirely new field which aims towards the improvement of tasks that musicians do, such as composing music, creating music (by playing an instrument), recording music, editing music, and playing back recordings. Music retrieval (searching music in large databases) has recently become a heavily researched field.

3.3 Culture
Speech processing usually relies on probabilistic methods. Feature extraction has become very unimportant, usually a set of standard features is used, such as: MFCC, PLP, RASTA, LPCC. These were developed in the early days of signal processing. Recently, prosodic features have caught the attention of scientist. Prosodic features are statistically invariant of what people have said and thus seem to be an indication of who speaks or his or her emotional state. Prosodic features include: Pitch, higher Formants, Long-Term Average Spectrum (LTAS), Harmonics-to-Noise Ratio, Speaking Rate (syllables per second). Machine learning techniques used for various speech tasks include: Gaussian Mixture Models (GMMs), Neural Networks also called Multi-Layer Perceptrons (NNs or MLPs), Support-Vector Machines (SVMs), and Hidden Markov Models (HMMs).

It seems that many tasks are always approached with an MFCC/GMM/HMM approach on the first try. This means, MFCC features are extracted from the audio tracks, GMMs are trained using expectation maximization and HMMs are used to model the time-dependencies between the frame-based GMM classification.

A very important aspect of the speech community is that any result must be benchmarked carefully using a publicly available dataset such as those from NIST or Linguistic Data Consortium (LDC). Papers wont be accepted otherwise. Error measures depend on the task: In ASR, Word Error Rate (WER) is used. Speaker recognition uses Equal Error Rate (EER), and diarization uses Diarization Error Rate (DER). Other measures such as Precision/Recall and F-Score, DET and ROC curves, confusion matrices, or simple false alarms/false positive percentages are used as well. One can say, that if a speech researcher does not know his current score on a public benchmark, he or she won’t be taken seriously among his or her colleagues.

Generally, for a paper to be accepted at a conference, showing that a method works is much more important then the originality of the idea.
3.4 Automatic Speech Recognition
Since ASR is the most important field in Speech Processing, here is an example of how an automatic speech recognizer works.

Speech recognition engines are usually very big systems. State-of-the-art speech recognition engines contain the work of many many PhD dissertations and are not a single-man effort. As a consequence, complete speech recognizers barely exists in universities. Universities usually only deal with certain aspects of the task. It is a domain of companies and research institutes. A natural consequence is that a speech recognition engine is divided into many small pieces that are glued together according to the task. Of course, there are training modules and test modules. Test actually stands for the actual speech recognition modules. The big picture is as follows:

- Feature Extraction
  The sampled audio data (wave file) is transformed into a different representation that allows an easier analysis. Mostly MFCCs are used.

- Speech Activity Detection
  The first thing to do after feature extraction is to get rid of any non-speech that there might be. For example coughs, laughter, door slams, music, etc... This is usually done using a trained approach.

- Feature Normalization
  After the features have been extracted and we got rid of all non-speech, one tries to make the features invariant to anything but the spoken words. Ideally, we want to eliminate any statistical dependency on the speaker or the channel (microphone, room reverberation). Therefore many techniques exist to normalize features, some are very basic, like Gaussianization, some of them are pretty advanced like Vocal Tract Length Normalization (VTLN).

- Recognition
  Now that we have audio that hopefully contains only speech and features that are invariant to everything but the actual spoken words we use GMMs/MLPs/SVMs to compare the spoken words on different levels (basically we use different window length) to the recorded and annotated words in our acoustic models. So we try to recognize "a" by comparing it to all instances of "a" stored in our acoustic model. It is considered to be an "a" if it is very close to all the other "a" and not so close to any other acoustic element, such as "e" or "o". Using different window length one can compare on sub-phoneme, phoneme, syllable and word level.

- Decoder
  Once small-scale recognition (e.g. phonemes, syllables, etc..) is performed, we have to glue the pieces together. This is done by using the language model. For each acoustic instance the recognizer usually outputs a set of alternatives with probabilities. So we have to make sure that individual phone combinations exist, syllables fit together to form words, and words actually form grammatical correct sentences. This is the hardest part.
The problem here is to choose the most likely combination of phonemes, syllables and words, according to the recognizer output. A very hard problem is to handle words that are not part of the language model.

- Textual Postprocessing
Once decoding is done, the output probably looks like this: "andhe said the quick brown fox jumped over the lazy dog named Bruno". So what one has to do now is take into account prosody, speech pauses, and other hints to detect sentence boundaries so that punctuation can be done. Also named entities should be detected so that capitalization works. Hopefully the output then looks like this: And he said: "The quick brown fox jumped over the lazy dog name Bruno".

4. Computer Vision
Computer Vision is actually a realm of disciplines. For the sake of simplicity, this document will divide it into two main categories: Image and video processing and image and video analysis. A third category is computer graphics, which deals with the display and rendering of images and videos. Like speech synthesis, this is a separate field and is not regularly using machine learning techniques.

4.1 Image and Video Processing
Image Processing is a rather traditional field. Many technologies in there are not considered machine learning techniques but simply math operations. Much of it is derived from the field of signal processing. The most important tools are: Fast Fourier Transform, Convolution with Kernels, and Morphologic Operations. Using these, an image can be blurred, denoised, edges can be detected, and so on. Other important operations include resizing and color correction. Image and video compression have dominated the field in the past decades. An overview is for example provided by:


4.2 Image and Video Analysis
Image and Video analysis deal with the handling of the content of images and videos. Main subfields are: image and video retrieval (finding all images that contain object x), image and video segmentation (finding the exact boundaries of image objects or video scenes), object recognition (detecting the particular objects, e.g. is there a face or a person in the image), and tracking (what is the location of a particular object).

Images and videos require a relatively high sampling, measured in dots per inch or pixels. Therefore it is very rare that images and videos are actually stored in an uncompressed way. Unlike speech algorithms, computer vision algorithms therefore have to be invariant against various compression artifacts, although they mostly work on uncompressed data. Like in speech processing (see above), this has several consequences:
a) Image and video processing, especially, if it is to be performed online and in realtime cannot rely on highly elaborated machine learning techniques. One hopes to find features that can be thresholded easily.

b) Scientific progress is considered fast-pace. A typical publication is 8-10 pages (in speech 4-6) double column.

c) Image and video processing is just starting to get a benchmarking culture.

d) The majority of the approaches seek to work online i.e., realtime and incrementally as new data comes in because there is a large range of consumer demand for image and video processing methods that are applied in editors.

4.3 Culture
Similar to speech processing, image and video analysis usually relies on probabilistic methods. Machine learning techniques used for various tasks include: Gaussian Mixture Models (GMMs), Neural Networks also called Multi-Layer Perceptrons (NNs or MLPs), Support-Vector Machines (SVMs), and Hidden Markov Models (HMMs). However, for many problems, non-probablistic methods have also shown to work, here distance metrics play a major role. Feature extraction is an important research part of every paper. Other than SIFT for image retrieval, there is actually no standardized or commonly used set of features, although 8x8-block DCT coefficients and optical flow (the set of all motion vectors) seem to be very predominant. Usually, the color space of an image or video is discussed, with the standard spaces being RGB, YUV, HSI (or HSV), and recently LAB. Edge detection (also called shape extraction) and color histograms are both rather simple and effective for various tasks and are therefore commonly used.

In image retrieval, common datasets are often used in order to make results comparable. Known datasets include: Corel Stock Photo Library or LabelMe by MIT CSAIL. Accuracy is usually measured in Precision, Recall, and F-measure (synonym for F-Score). NIST provides a set of tasks and a dataset that is evaluated regularly under TrecVid. The Clear evaluation was also initiated by NIST. Other than those, many benchmarks and datasets exists created by individual institutions or researchers (e.g. the Berkeley Image Segmentation Dataset and Benchmark).

Generally, for a paper to be accepted at a conference, the originality of the idea is more important then extensive benchmarking. Anecdotical evidence is often enough.

4.4 Image Segmentation
Object extraction from images and videos (interactive or non-interactive image and video segmentation) is an important field and in computer vision. Here is an example of how a semi-automatic object extractor in GIMP (SIOX) works.

- User Interaction
Given an image, a free-hand selection tool is used to specify the region of interest. It must contain all foreground objects to extract and as few background as possible. The pixels outside the region of interest form the sure background while the inner region define a superset of the foreground, i.e. the unknown region. A so-called foreground
brush is then used to mark representative foreground regions. The algorithm outputs a selection mask. The selection can be refined by either adding further foreground markings or by adding background markings using the background brush.

- Feature Extraction
The algorithm then converts all pixel into CIELAB space.

- Model building
A set of representative colors for sure foreground and sure background, the so-called color signatures, are created by a clustering technique.

- Classification
All image pixels are then assigned to foreground or background by a weighted nearest neighbor search in the color signatures.

- Postprocessing
Standard image processing operations like erode, dilate, and blur are applied to remove artifacts and the largest connected foreground component is found.