EE3F1 Multimodal Interaction

Coursework Assignment

18/11/12

Yousef Amar (1095307)

Contents

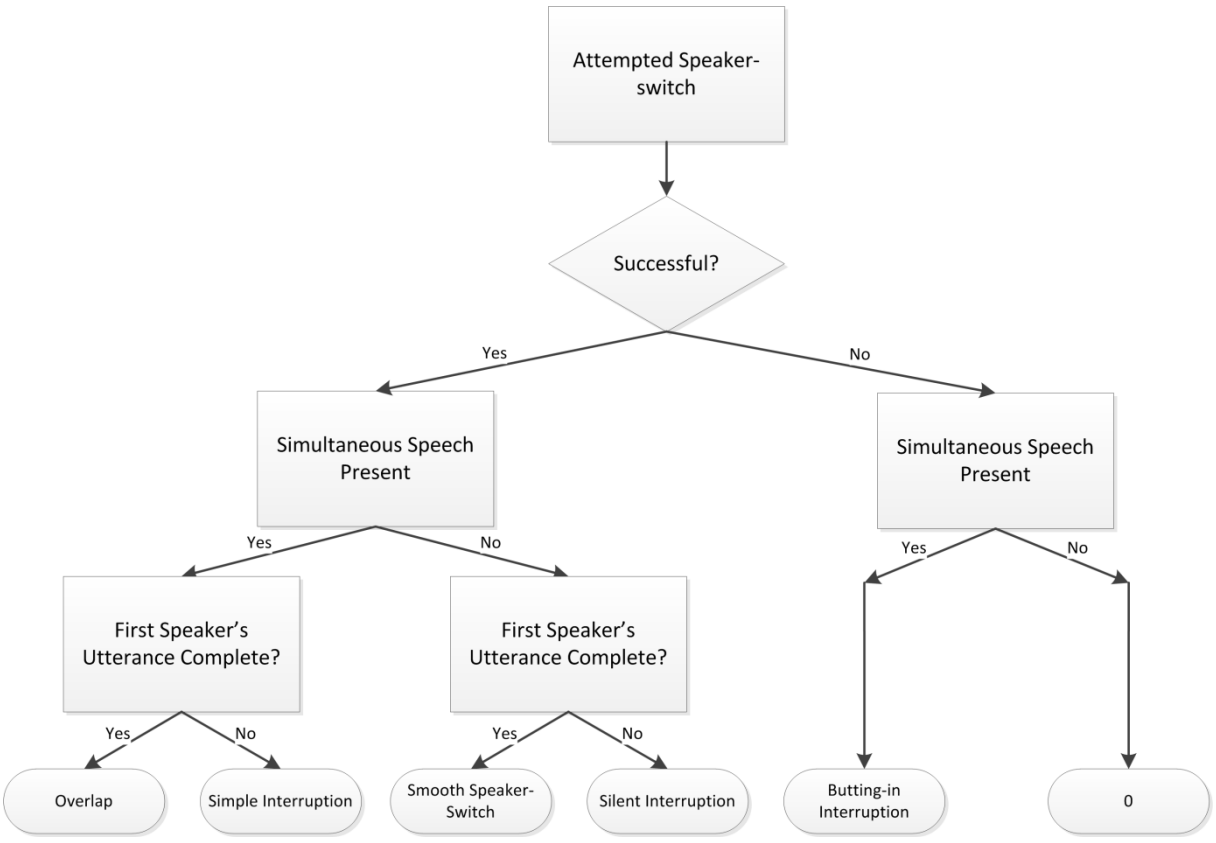
Introduction

When studying human social interaction, turn taking behaviour is not deeply significant yet remains, to this day, difficult to automatically predict, let alone model, as humans themselves have issues with classification especially seeing as we live in a time of globalisation where people can have a myriad of habits and display heavily contrasting cues.

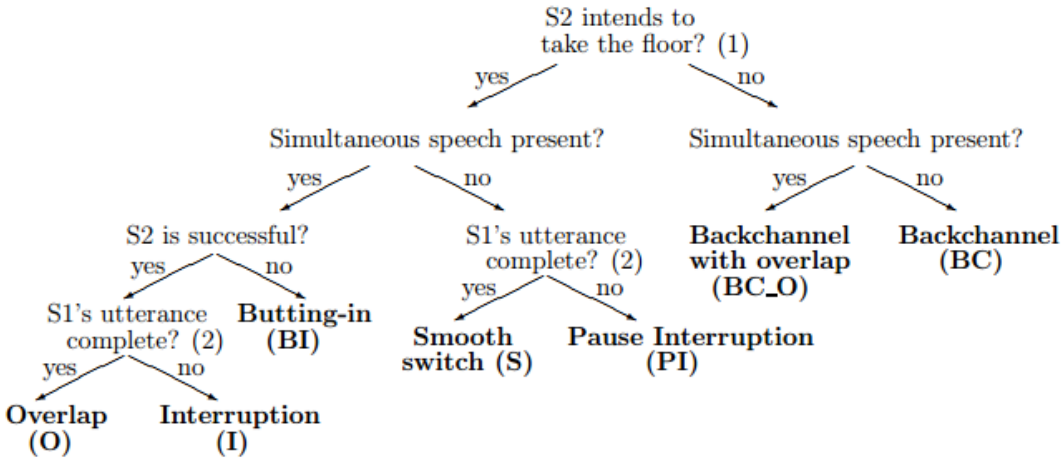
Using information gathered from the video and audio data provided for this assignment from the AMI Meeting Corpus, and other research within the field, a realistic classifier can be designed that predicts which person speaks next.

Literature Review

The first thing one would have to do is define “turn-taking”; what counts as a turn? The most basic definition would be any vocalisation above a certain threshold (Jaffe & Feldstein, 1970) and a more practical definition could be any utterance greater than 5 seconds (Kendon, 1967). (Beattie, 1982) however, split speech into several categories.



This has since been further refined into turn "tiers" (Hirschberg et al., 2004).



Most tiers are relatively self-explanatory with the exception of backchannel cues which are explained further down. For the purpose of this assignment, a turn is defined as being "taken" on Overlap, Simple Interruption, any Interruption (provided the speaker allows the interruption which can be predicted on the basis of speaker dominance and turn frequency), and Smooth Switch.

Turn-taking behaviour can be categorised into four main areas (Grün, 1998): turn-yielding, turn-requesting, turn-maintaining, and backchannel cues. The following are explanations of cues that are monitored in the video analysis and taken account of in the classifier design based on observation and literature: (Duncan, 1972); (Wiemann & Knapp, 1975); (Malandro et al., 1989).

Turn Yielding

- Intonation drift
  - "Pitch-level termination junction" (Duncan, 1972) ≠ 2 2| at the end of a phonemic clause.
  - Following (Trager & Smith, 1957) notation, the 2s in "2 2|" refer to intermediate pitch levels (where 3 is high and 1 is low) and the "|" refers to sustention.
- Drawl on final or stressed syllable of a terminal clause
- Sociocentric Sequences
  - For example "but uh", "or something", "you know" or other filler phrases
- Paralanguage (Pitch/Volume) drop with sociocentric sequences independent of terminal clause
- Significant pauses or lack of **turn-maintaining** cues ("um", "ah") or volume increase on attempted interjection
- Grammatical syntax (not considered due to complexity)
- Gesture Termination/Relaxation (though not all people speak with their hands)

Turn Requesting

(In the absence of these, it is difficult to predict who will speak next)

- Opening of the mouth while a speaker is talking or finishing
- Change of posture (leaning in) and gesturing or tensing
- Stuttering (Malandro et al., 1989) or buffering before speaking
    - Buffers (Wiemann & Knapp, 1975) can be defined as short interjections in pauses to get a turn (e.g. “but uh”, “you know”)

Determining the Next Speaker in Absence of Turn-requesting Cues

- If the chair controls who speaks verbally and/or aided through beckoning (not applicable)
- If a turn yield is significantly prominent, the gaze direction can be used
- Speakers that display **back-channel** cues can be eliminated
    - Back-channel cues are utterances that do not count as turns as they are usually just used to signify attention (e.g. “mhm”, “oh”, nodding) or express the rejection of a turn-yield or unwillingness to take over.

The modality that undoubtedly has the most significant role is speech (words and intonation) since even as humans, it is what we use to recognize when people are done talking or want to talk. This is closely followed by gesturing. While body language may be very important in communication, it is less important in turn taking as it can be ambiguous and inconsistent. It would also be difficult to sense and analyse. Gaze, specifically the motion of the head rather than eye movement or a combination of the two, can be broadly indicative of predicting the next speaker, but also requires extra processing.

Video Analysis

Based on previously discussed research and observations within the video and in real life, information is gathered from the video. The methodology for counting the frequencies of occurrences involves watching the videos twice, the first time for general observation and the second for more details, and tallying cues, only pausing to take more detailed notes. This provides more meaningful data compared to an exhaustive chronology of all cues. It is important to remember that values gathered in this manner may not be error-free as it is easy to miss cues when watching four people at once such a limited number of times. It does however give useful estimations that can be help design a working classification system.

For utility, the speakers are each assigned a number. The apparent chair (off-camera) is 0 while the other speakers are numbered 1 to 4 respectively in the order they are shown in the SMIL media file. Unfortunately the data itself is difficult to work with; not only are the individual videos quite low resolution and significantly shorter than the matching audio (about 30 minutes or ~1/3 of the entire meeting), but they show pretty much only the faces which is unfavourable as this implies that bodily gestures, cues most dominant in human conversation, cannot be taken full advantage of in the turn prediction process. Furthermore the audio is somewhat inconsistent and quiet especially considering that some of the speakers are quite introverted and prone to mumbling.

Additionally, notes were made that could come in useful in designing the classifier. The following information was obtained.

Relevant and Noticeable Turn Taking Cues

Cues are split into, general (top), turn yielding (middle), and turn requesting (bottom) while turn maintaining and backchannel cues are any talking that is not classified as yielding or requesting.

Speaker:	0	1	2	3	4
Turns Taken	108	44	29	26	79
Apparent Turns Yielded	33	12	4	2	21
Apparent Turns Requested	34	26	8	7	31
Intonation ≠ 2 2|	48	19	8	13	40
Drawl	26	4	3	5	20
Sociocentric Sequences	1	1	0	0	0
Paralanguage	4	2	0	1	0
Significant Pauses	29	4	3	1	16
Grammatical Termination	48	15	8	10	32
Gesture Termination/Relaxation	N/A	0	0	0	0
Stutter/Buffer	23	9	4	3	21
Opening Mouth before taking turn	N/A	8	1	2	8
Change of Posture on request	N/A	2	1	0	1

Notable Occurrences

Chronological

- When all speakers display back-channel behaviour, it is almost guaranteed that the turn will default back to the last, and usually most dominant, speaker.
- After a question, the asker's voice tone increases in pitch and their gaze is aimed at the person the question is meant for. Chances are that person will speak next. This variation in intonation changes is accounted for (≠ 2 2|).
- Non-native speakers (usually with an accent), for example 0, 1, and 2 seemingly, may explain the lack of sociocentric sequences as well as differing paralanguage and intonation that what is rooted in British/American English.
- Pauses without grammatical termination are ignored by the others yet usually are not in time with turn yielding cues.
- Immediate example of back-channel behaviour at 2:15 is speaker 3 nodding and then the speaker remains. As the speakers all seem relatively passive, the turn is constantly defaulting back to the chair (unseen) as they display back-channel cues making the data quite suboptimal for analysis.
- Approval ("yeah", "mhm", BC to chair) is mostly done by the speaker 0 (chair), the most dominant.
- Sound cuts off at the 3:20 mark to 3:30.
- 7:12, 13:13, 14:47: Speaker 3 yawns that could be mistaken for getting ready to speak, itself a BC cue (grooming).
- 14:01: Speaker 1 significantly prepares to speak for a long time opening her mouth and looking for a pause to switch into.
- 59:52: All videos go off but by now their voices can be distinguished from one another.
- The second last different person to talk is always more likely to talk again in a one on one manner in a small meeting like this.

General

- Gaze is not often directed to the next speaker at yield; it may be insignificant in this type of conversation, possibly requiring more people.
- The videos do not show the speakers' hand which disallows many useful observations.
- The video data would be difficult to classify; most speakers are looking away when talking, some speakers are way too quiet, the resolution is too low to see any sort of emotion or detect mouth movement accurately, and their seating arrangement is sub-optimal.

Classifier Design

Using the information gathered in research and analysis, a classification system can be designed. The classifier design is created with general applications in mind rather than specifically designing it to suit the videos analysed. This is because further data can be considerably valuable and should not be ignored in the design just because it is unavailable in the data provided. This includes hand gestures, facial expressions, gaze direction and the like.

Relevant Multimodal Behaviours

From the videos, certain behaviours were more prominent than others in inferring turn taking cues. At the same time, some may not be very indicative or particularly useful in designing the classifier but are easy to extract and process and help the classification. The multimodal behaviours and their respective modalities that are relevant to the design, disregarding those that will not be considered, are as follows in order of practicality and feasibility.

Turn Yielding

Behaviour	Modality/Modalities
Change in intonation	Speech
(Drawl)	Speech
Significant pauses	Speech (or lack thereof)
Gesture termination/relaxation	Gesture
Sociocentric sequences	Speech
Paralanguage	Speech
(Grammatical termination)	Speech

Turn Requesting and Prediction

Behaviour	Modality/Modalities
Stutter/Buffer	Speech
Opening mouth	Gesture (possibly even emotion)
Gaze Direction	Gesture

Turn requesting cues are in order of accuracy over time; the top behaviours are useful for predicting a turn request right before or as it is happening while the bottom can make long term predictions. Either way the prediction becomes more and more accurate exponentially as the event draws nearer.

Feature Extraction

Certain features matching the multimodal behaviours need to be extracted in order for the information gathered to be of use. Realistically, it should be done in a non-intrusive manner (i.e. no coloured markers on your person or anything that should hinder the process or make the analysis of large quantities difficult) and ideally be using just a decent microphone on every speaker as well as a camera with a decent resolution aimed right at them from the front. Speakers should not have to turn their heads more than 90° in either direction in order to see all other speakers. The reason behind this and why a single microphone should not be used is discussed later.

Realistically, features needed for classification could include:

- Pitch
- Volume (for paralanguage; unless differences are minimal)
- Waveform of an utterance (with limited classes)
- Gaze direction
- Upper body movement amount
- Mouth area coverage

Feature extraction methods for speech would simply be measuring the frequency and amplitude of and audio signal to later differentiate and see the change in pitch or volume over time. This is not dissimilar to part of classifying eye movement (fixations and saccades). As for isolating utterances, the volume data can be used to identify pauses, simply by setting a threshold rather than doing complex classification, and separating segments of speech using pauses that are longer than a certain amount of time as separators. A value for the threshold time can either be set or can become a variable in the system that is continuously refined as the speakers talk by measuring the new values and correcting (e.g. using Bayes theorem or a Kalman filter).

When it comes to video, accuracy is not the first priority as it only supplements the classification rather than build it. Simple image processing can be used to detect if a speaker's mouth is open and check that against if they are actually saying anything at the same time. This can be as simple as looking at how many pixels are beyond certain darkness in the mouth area that would be based up how many actually are while they are talking. Once again a complex classifier can be designed for this purpose but would be overkill.

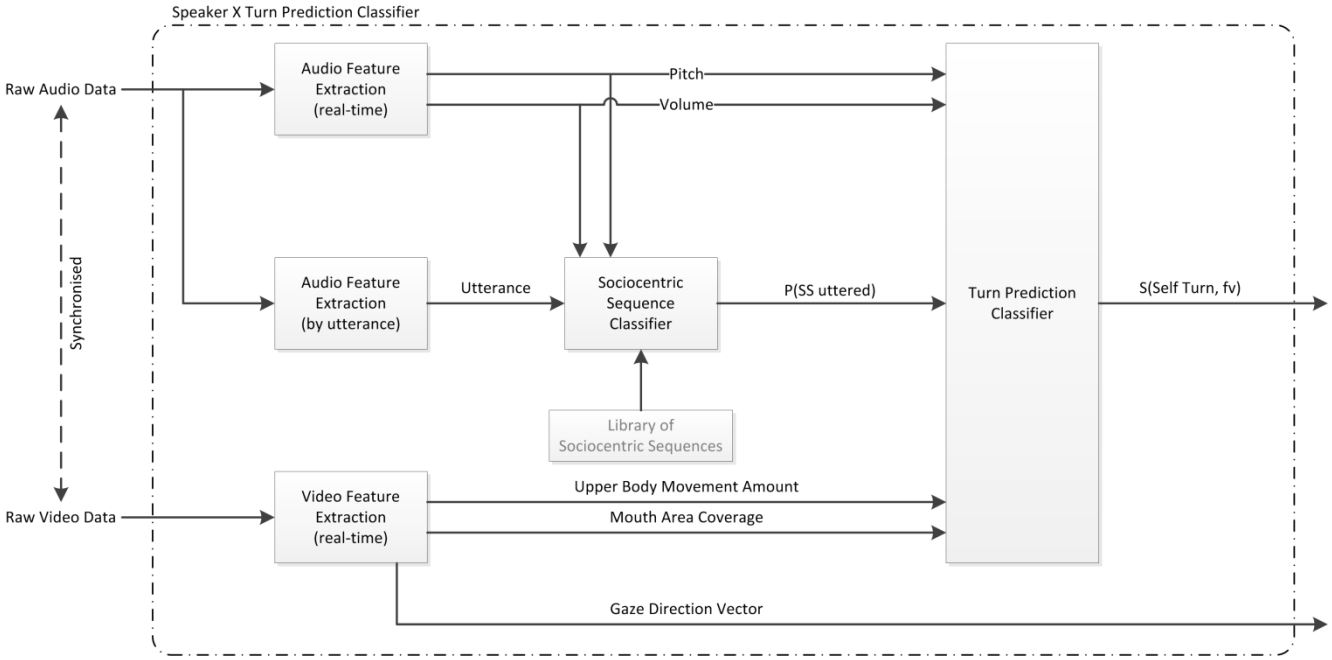
Likewise, gesturing or other upper body movement does not need to be tracked carefully as the result would not have a deep impact. As with mouth movement, another threshold can be set for how much motion can be detected around the speaker's upper body, an image processing problem, and give a binary true or false on if the speaker is gesturing with their hands and arms.

Gaze direction is a whole different matter, yet fortunately speakers turn their entire heads towards another person in conversation instead of just their eyes or a combination of the two. Participants probably would not want a gyroscope fitted to their heads and doing that could actually have an adverse effect as it might make them self-conscious about their gaze, almost like Observer effect in physics that alone the act of measuring something changes the measurement. Measuring the distances between both eyes and the mouth and doing simple math may be enough to estimate the gaze direction and would probably be sufficient and feasible since even digital cameras can do it.

Classifier Architecture

All the points previously discussed can used to abstract a top-level architecture design. This design is split up into multiple parts with some classifiers having their own detailed internal designs which are explained in detail later.

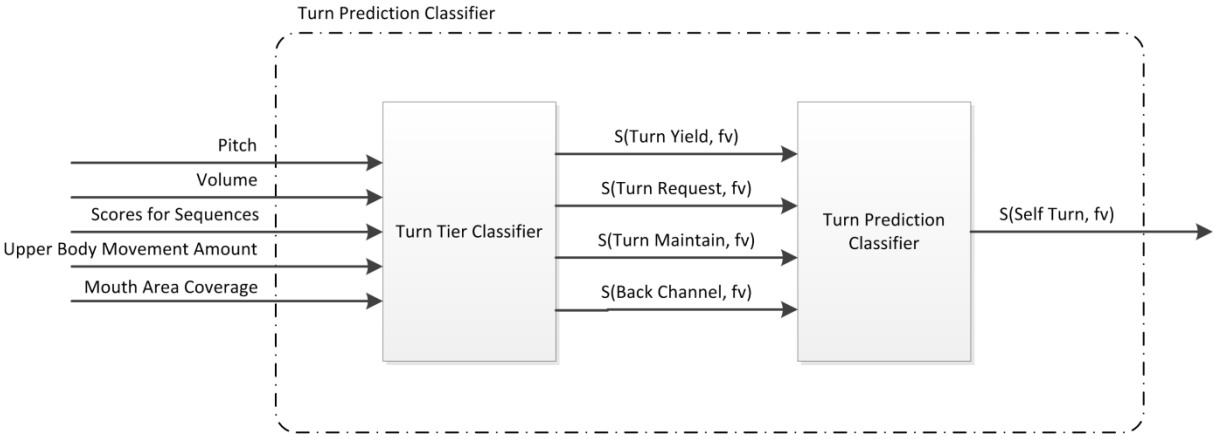
Speaker X Turn Prediction Classifier



The function of this sub-architecture is mostly self-explanatory; for a single speaker, with isolated video and audio data, a probability that they will continue talking is derived together with their gaze direction. The raw audio data is synchronised from the start such that each feature vector for audio data has corresponding video feature vectors. If the sampling rates differ, simple linear interpolation would be sufficient to match the data up. A mixture of early and intermediate integration is used but ultimately, a single joint classifier would give the final results so no classwise/cross-class fusion is needed but some intermediate feature fusion. Furthermore dimension reduction would only result in the loss of information.

Once the feature vectors are extracted using the methods previously mentioned, the sociocentric sequence classifier outputs the probability that a sociocentric sequence has been uttered. Of course speech classifiers are in themselves complicated systems, but in this case, phonemes can be completely ignored and instead entire phrases can be compared to the incoming signal. Rather than creating scores for specific sequences, that may differ based language, whether the speaker is a native speaker, and other factors, a single one-dimensional vector can be obtained that indicated the probability that a sociocentric sequence has been uttered. The classification would still be generative as one is finding the differences between the uttered speech waveform and the model data and then deriving the probability. **Note that feature fusion through feature vector concatenation is implied and an explicit feature fusion block is not depicted for clarity.**

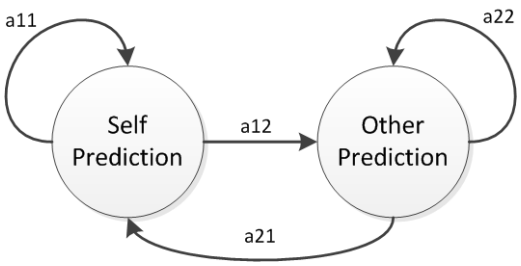
Turn Prediction Classifier



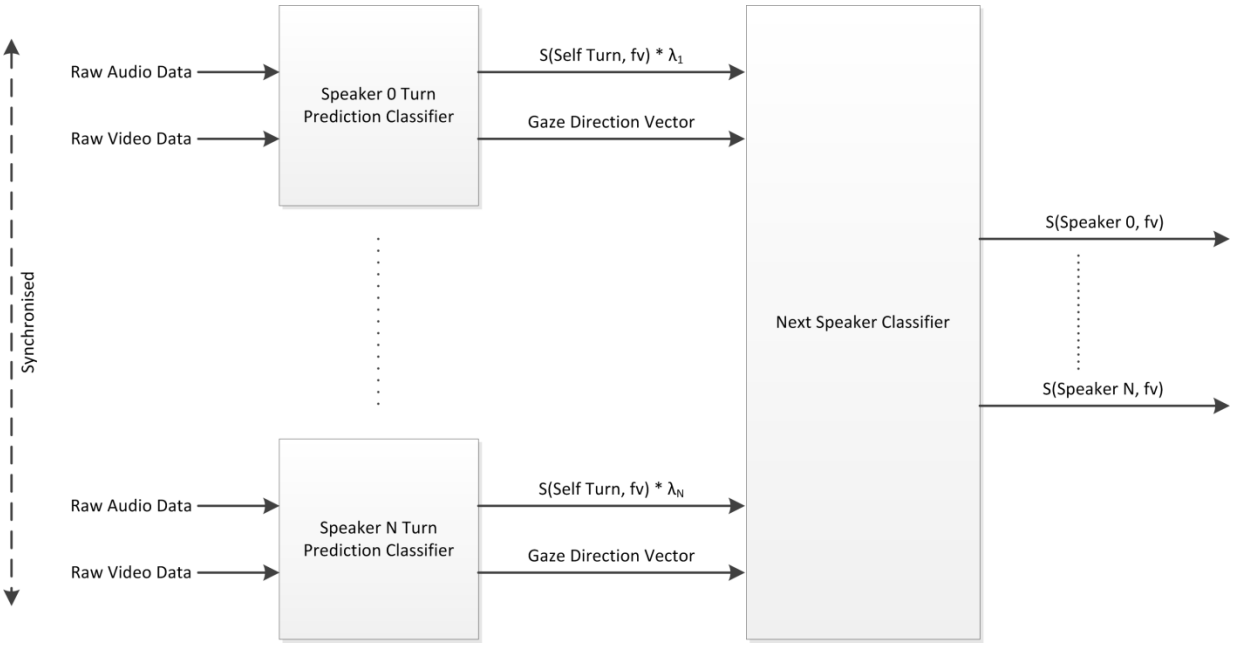
The turn prediction classifier would internally calculate the scores for turn yielding, requesting, maintaining, or backchannel cues. This is done in a sequential, control flow manner as there is no real algorithm that can be used. The amount of change in pitch and volume can be directly mapped to the probability that they correspond to each class as a function and all the information is combined by averaging all values after they have been normalised potentially with weightings that modify the confidence that each modality or specific feature would yield accurate results.

The probability that the next turn is the same speaker (i.e. they are in the middle of their turn) would then become a function of the scores from the prior classifier also just sequential control flow, averaging, and basic maths.

An ergodic Hidden Markov Model with the states “self prediction” and “other prediction” could help streamline the classifier even further through setting the 2D state transition probability matrix dynamically based on the relative amount of time that particular speaker speaks in relation other speaker. For instance, if the spoke 80% of the time in 30 minutes, it becomes more likely that they sustain and gain turns but if that value is reduced to 40% in the hour following that, the speaker becomes less and less likely to do so.



Complete Classifier

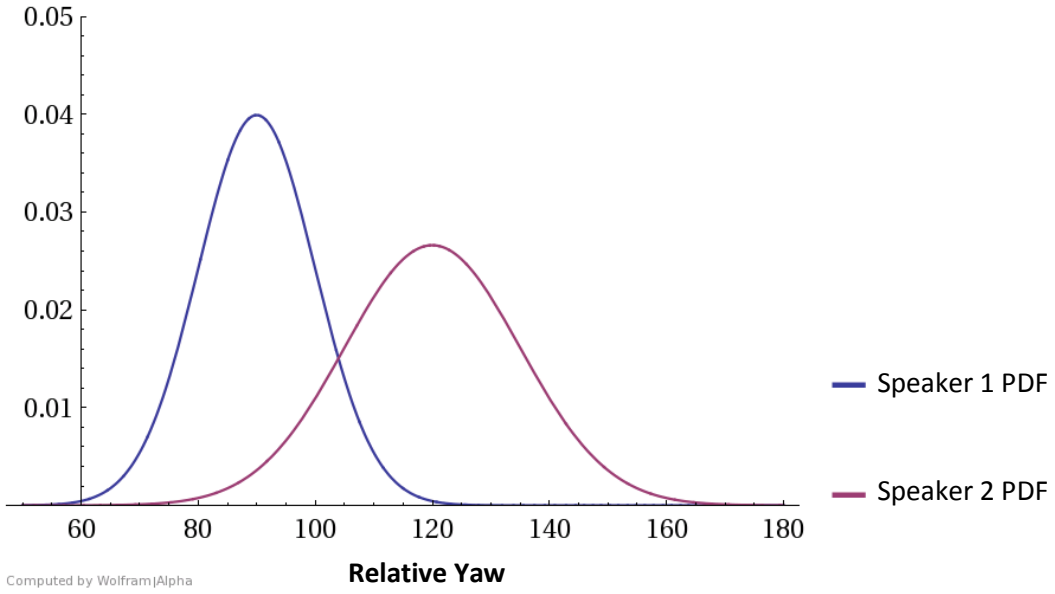


Finally, a multiplicity of speakers, each with their own turn prediction classifier, are combined in a modular fashion. One minus the probability that the next turn will be themselves gives the probability that it will not, effectively compacting a 2D feature vector into one dimension as the values correlate directly.

For gaze direction, only one value is really needed: a horizontal angle (around the “up” axis) relative to north or a fixed and common direction. The pitch of the gaze can be discarded completely and only the yaw for each speaker’s gaze direction is required. Then, provided that the speaker is displaying turn yielding or back channel behaviour by having a lower probability of keeping the turn, one is effectively building a Gaussian Mixture Model.

Prior to the meeting, several Gaussian distributions can be set as models of which the x-axis is yaw in degrees. Each speaker would have a different set of distributions relative to where the other speakers are. The mean, ideally the angle they would look to see the speaker dead-on, would change depending on where the speakers are sitting in relation to each other, while the variance would depend on the distance from each other probably. An example for speaker 0 is given below. Note that the curves do not relate to the videos analysed in any way and were plotted for the purpose of explanation. In the case of the videos analysed, the positions of the speakers could be deduced from the top-down camera but that is beyond the scope of this assignment.

Example Model Data for Speaker 0



In this instance, speaker 1 is sitting at 90° relative to speaker 0 (the angle of reference being the opposite direction their camera is facing, i.e. when their gaze is perpendicular to the camera) and speaker 2 at 120°. Speaker 1 and speaker 2's distributions have the unrealistic variances of 10 and 15 respectively for illustration purposes. The difference in variance could correspond to how close somebody is or maybe they are moving around for some reason.

It is important to note as a caveat that these are but estimations; several assumptions are being made. The first is that is it even suitable to model as normal distributions. The distribution is not necessarily symmetric. For example, a listener might be less inclined to turn their head all the way if a speaker is sitting at an uncomfortable angle. In that case the curve might stretch further to the right as you move away from 0° down and further to the left as you move from 0° up. It is however a decent enough estimation and works better than the alternatives.

It is clear to see however that the gaze angle of speaker 0 is a mixture of the distributions, with weightings dependant on how much speakers 1 and 2 talk relative to each other and speaker 0, and finding out which curve, and thus speaker, it is most likely to fall under mathematically using the GMM and PDF formulae. Ultimately, one can deduce who the speaker in question is looking at and thus who might take the next turn.

The models can also be built during the meeting in real-time as we know the person who's turn it is will probably have all gaze directed towards them so you could infer their position, or angle relative to all of the listeners, just from the gaze of the listeners or possibly even using the Monte Carlo method. This would be more complicated however and while "training", the classifier would be prone to inaccuracies.

Additionally, the second last person to speak is more likely to speak again than usual since they could be involved in a momentary one-on-one conversation. To account for this, the second last speaker is remembered and biased in calculation; a further generative model (HMM/DBN). So not only is the original outputs for each speaker given a weighted average modification based on their dominance in the conversation up to any given point, the weightings are effectively modified themselves to allow the second last speaker to be positively biased.

Finally, once the scores are calculated, the highest one is selected as the prediction. This is a single final score that has not undergone any sort of fusion. The result however is fed back into the classifier to use for subsequent classification.

Variables needed in real-time

- Relative amount of time speaking for any given speaker
- Number of turns taken for any given speaker
- Current speaker (audio + predicted turn + gaze)
- Predicted speaker
- Time for a prediction to be fulfilled or be unsuccessful (predicted + actual)

Conclusion

Subject to implementation and testing, the design should be able to predict the speaker immediately following the current speaker, in the worst case at a second prior to termination, reasonably well. Any earlier than that and the prediction becomes tentative. This design would also solve issues that arise in the disordering prosodic turn-taking cues (Cutler & Pearson, 1985). Through combining multiple modalities, this multimodal system, like so many others, becomes greater than the sum of its parts.

Bibliography

Beattie, G. W. (1982). *Turn-Taking and Interruption in Political Interviews.*

http://www.cs.columbia.edu/~sbenus/Teaching/APTD/Beattie_1982.pdf

Cutler, A., & Pearson, M. (1985). *On the Analysis of Prosodic Turn-taking Cues.*

http://pubman.mpdl.mpg.de/pubman/item/escidoc:76883:7/component/escidoc:506929/Cutler_1985_On%20the%20analysis.pdf

Duncan, S. (1972). *Some Signals and Rules for Taking Speaking Turns in Conversations.*

http://hhhuang.homelinux.com/projects/GECA/papers/Duncan1972.pdf

Grün, U. (1998). *Visualization of Gestures in Conversational Turn-Taking-Situations.*

http://coral.lili.uni-bielefeld.de/Classes/Winter97/PhonMM/UlrichGruen/cues.htm

Hirschberg, J et al. (2004). *Turn-taking Labeling Guidelines.*

http://www1.cs.columbia.edu/~agus/games-corpus/guidelines-turn-taking.pdf

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue.*

Kendon, A. (1967). *Acta Psychologica.*

Malandro et al. (1989). *Nonverbal Communication.*

Trager, G. L., & Smith, H. L. (1957). *An outline of English structure.*

Wiemann, & Knapp. (1975). *Turn-Taking in Conversations.*