Identifying Radio Communication Inefficiency to Improve Air Combat Training Debriefings

Hanna Lilja, Joel Brynielsson, Sinna Lindquist
FOI Swedish Defence Research Agency
SE-164 90 Stockholm, Sweden
hanna.lilja@foi.se, joel.brynielsson@foi.se, sinna.lindquist@foi.se

ABSTRACT

In flight simulator training, the retrospective debriefing activity is the essential tool for reflecting upon the conducted exercise. The debriefing typically serves to highlight carefully chosen exercise moments that can help the trainees to gain further insight. These moments are revisited from a number of different perspectives in order to make the overall picture of what actually happened during the exercise clear to the trainees. It is therefore of utmost importance to be able to identify the most instructive moments, and compile the data to be used for the debriefing. This is a delicate and critical task that needs to be undertaken swiftly so that the debriefing activity can be conducted directly after the flight exercise. This paper presents a methodology and prototype system for identifying moments of inefficient radio communication as a means to obtaining additional data for improving many versus many air combat flight simulator training debriefings. The work fits into an overall systems perspective where off the shelf speech recognition technology is used to obtain textual representations of the radio communication call sentences, which are then fed to a tailor-made machine learning classifier that is capable of detecting incorrect terminology/phrasing, unnecessary repetitions, unnecessary stalling of the radio communication channel, etc. With this overall systems perspective in mind, radio communication recordings originating from ordinary exercises being held at the Swedish Air Force Combat Simulation Centre have been used to construct classifiers capable of capturing communication patterns that are interesting to look at from an efficiency point of view. To validate its usefulness, the developed classifiers have been used to create visualizations that help identifying additional exercise moments in terms of inefficient radio communication that ought to be highlighted alongside the ordinary 3D visualization of the air flight, the cockpit displays, etc.

ABOUT THE AUTHORS

Hanna Lilja is a research assistant at the Swedish Defence Research Agency (FOI) in Stockholm, Sweden. Hanna holds an M.Sc. in Engineering Physics from the Royal Institute of Technology (KTH) in Stockholm, Sweden. Her research interests include machine learning, natural language processing, and visualization, with a particular interest in flight simulator applications.

Joel Brynielsson is a deputy research director at the Swedish Defence Research Agency (FOI) and an associate professor at the Royal Institute of Technology (KTH). Joel is Docent (Habilitation) in Computer Science (2015), and holds a Ph.D. in Computer Science (2006) and an M.Sc. in Computer Science and Engineering (2000) from KTH. His research interests include uncertainty management, information fusion, probabilistic expert systems, the theory and practice of decision-making, game theory, web mining, privacy-preserving data mining, cyber security, and computer security education.

Sinna Lindquist is a senior scientist at the Swedish Defence Research Agency (FOI), and holds a Ph.D. in Human-Computer Interaction from the Royal Institute of Technology (KTH) in Stockholm, Sweden. Her research interests include pedagogical issues regarding decision training, decision support systems, crisis management, usability, as well as methodological issues regarding field studies and user-centered design activities.

Identifying Radio Communication Inefficiency to Improve Air Combat Training Debriefings

Hanna Lilja, Joel Brynielsson, Sinna Lindquist
FOI Swedish Defence Research Agency
SE-164 90 Stockholm, Sweden
hanna.lilja@foi.se, joel.brynielsson@foi.se, sinna.lindquist@foi.se

INTRODUCTION

In flight simulator training, the retrospective debriefing is an essential tool used to enhance the learning experience. The debriefing of a many versus many air combat training simulation is used to evaluate the performance of individuals as well as that of the team. Efficient spoken communication is an essential part of any cooperative work, and no less so in the case of radio communication during air combat. However, there is currently a lack of tools to support evaluation of this type of communication. Without such tools the evaluation of this aspect of the simulation relies on an experienced instructor's ability to notice, remember and later point out instances of less-than-optimal communication as it is too time-consuming to replay the entire scenario in real time at every debriefing to manually do this analysis afterwards. It is not obvious what such tools should look like, nor is it obvious where the line between efficient and inefficient communication should be drawn. An important part of this work is therefore to understand how to represent the communication data in order to identify parts of the communication which are in some way inefficient.

This study was initiated by the Swedish Air Force Combat Simulation Centre (FLSC), which has provided the necessary data and required expertise to carry out the project. FLSC is a simulation facility for manned many versus many air combat assignments. FLSC offers a simulation environment for implementing scenario simulations for training, development, and practice that is adapted to operational requirements, as well as research regarding training, evaluation and flight simulation development. This also includes research for the development of measurement tools, and design support for the realization of technical concepts with a focus on measurement and evaluation of the tactical/operational capacity.

It would be beneficial to include the communication aspect in air combat debriefing in a consistent way, so that the general efficiency of the communication can be evaluated. The study reported on herein serves as a first step towards providing those performing air combat debriefings with the tools needed in order to make such an evaluation easier.

THE AIR COMBAT DOMAIN

In this section selected aspects of the application domain are described. Some parts are relevant in order to understand the problem, some to explain the methodology, and others are mainly of importance for the interpretation and discussion of the results.

Beyond Visual Range Combat

Beyond visual range (BVR) combat is, as the name suggests, combat where a pilot does not normally have direct visual contact with the enemy. Instead of eyesight, radar is used to locate threats and to guide missiles to their intended targets. In order to launch a missile a pilot locks their radar on the intended target. Information about radar locks and launched missiles show up on displays in the aircrafts of the other members of the group. The aircrafts are also equipped with a radar warning receiver (RWR), which alerts the pilot of incoming radar lock ons.

Two pilots working together as a pair is called a *2-ship*. A common setup is to put two such pairs together to cooperate in a group of four, a *4-ship*. The pilot in charge of such a group is the *4-ship lead*. Additionally, the group is supported by a *Fighter Controller* (FC) on the ground. In this setting a pilot gets information from their own radar, from the other aircrafts in the group (via data link), as well as verbally from the FC and the other pilots.

Radio Communication

The verbal communication between pilots and the FC is transmitted via radio. When training at FLSC there are usually two radio channels in use. The main channel is the one which both the FC and the pilots listen to and can speak on, while the other channel is used only internally by the pilots. In order to send a transmission on a channel, i.e., speak on the radio, the speaker has to press a button. This is referred to as push-to-talk (PTT). On a real radio channel anyone can transmit at any time. During simulation at FLSC, however, only one person at a time can speak on each channel. This means that whoever pushes their button first gets to speak and not until that person releases their button will anyone else be able to transmit.

It is not obvious how to define communication efficiency in this particular domain. The words and phrases used are very different from many other language applications, the grammar of an utterance does not necessarily follow normal language rules. The importance of time is also rather extreme, in comparison with other domains. In a survey among pilots of the Finnish Defence Forces some of the most common communication problems reported were multiple speakers transmitting at the same time causing overlapping speech, and missing acknowledgment calls. The former was considered a problem in particular by the fighter pilots. Additionally, 59% of the fighter pilots reported that most of the communication problems occurred during air combat exercises (Lahtinen et al., 2010). A study of speech acts during missile launch situations in simulated BVR combat scenarios has shown relations between speech acts, communicative problems and outcomes with respect to missile hit or miss, as well as overall mission success (Svensson and Andersson, 2006). The speech acts were manually coded into seven categories, and the communication was analyzed in three steps: identifying the communicative problems, localizing the types of situations where the problems arise, and finally determining the cause of the problem.

Terminology and Phraseology

The domain specific terminology and phraseology are described in a manual issued by the Swedish Armed Forces (2007). This manual also contains instructions for how and when to use specific terms and phrases. Only some basic concepts and terminology examples, relevant to BVR combat, will be presented here.

An important process during air combat, intended to maximize the situation awareness of the pilots, is *picture* building. The picture is a description of the air situation in terms of enemy units and unknown units. A *group* is a number of aircrafts within a certain distance of each other. Multiple groups with larger separation is described as a *package*. A group is described by number of contacts, formation, position and identity.

A formation is described by type (e.g., *box*, *wall*) and size (how *wide* and/or *deep* it is). The position of a group can be described with bearing and range in relation to a specified geographic point referred to as *bullseye*, with the addition of altitude. It can also be described in *BRAA* format. BRAA means bearing, range, altitude and aspect in relation to a fighter. A specific contact within a group is referred to with its direction (e.g., *west*, *east*) and depth position (*lead*, *middle* or *trail*). Aspect is used to describe the threat angle relative to own position (*hot*, *flank*, *beam* or *drag*) whereas *head on*, *crank*, *notch* and *pump* are used by pilots describing their own maneuvering. In addition to this, *track* (and direction) is used to describe the direction of flight of a non-friendly aircraft.

A *threat call* is a warning from the FC regarding an untargeted group which fulfills the range and aspect criteria of being a threat. A *spike call* is a pilot informing the rest of the group and the FC about a radar threat detected by RWR. A pilot may also request a *declaration* of the identity of a group or contact from the FC.

THEORY

The first part of this section concerns the field of machine learning. The second part describes feature handling, which can be considered a part of the machine learning process but is discussed here separately due to its importance to the problem discussed in this paper. Some relevant aspects of natural language processing are then presented briefly.

Machine Learning

Machine learning is a way of inferring new knowledge from data. It is useful when there is no algorithm available to directly perform a task, but when there are examples or past experience of the outcome of the task in the form of data.

In general, a model is constructed based on training data (the learning process), while the performance of the model is evaluated on test data that has not been used during the training. The model can be predictive or descriptive; the former is used to make predictions based on past experience, the latter to gain new knowledge (Alpaydin, 2006).

In supervised machine learning one wants to predict the value of an output variable, based on the values of the input variables. For such learning to be possible there has to be a large enough set of data instances for which the values of both the input variables and the output variables are already known. A typical supervised learning task is the classification task. In this case the output variable is typically discrete, denoting the class a data instance belongs to—a class label. Some classification algorithms deal only with two classes, i.e., a binary decision, whilst other algorithms can handle multiple classes.

The idea of supervised learning is to learn a mapping from input to output. The usual approach is to define a parametric model and then use a set of training data to learn the optimal values of the parameters by, for example, minimizing the approximation error or maximizing the classification accuracy. Evaluation of the performance of a classifier can be done statistically using an additional set of data for which the class of each instance is known. The class label assigned by the classifier is compared to the known true class and statistical measures of how often a new data instance is correctly classified can be obtained. One such measure is precision, which for a given class A is defined as

$$\text{precision}_A = \frac{TP}{TP + FP}, \qquad (1)$$

where TP is *true positives* (the number of instances classified as A for which the true class is indeed A) and FP is *false positives* (the number of instances classified as A for which the true class is not A). Another commonly used measure is recall, defined as

$$\text{recall}_A = \frac{TP}{TP + FN}, \qquad (2)$$

where FN is *false negatives*, i.e., the number of instances which belong to class A but were not classified as A.

Supervised learning is used in this paper as a tool for classification of utterances. This is a non-binary task, i.e., it involves multiple classes. The performance of these classification models are evaluated in terms of precision and recall.

Features

A feature refers to an element of the representation of the data that is eventually presented to a learning algorithm. Feature engineering is the process of constructing relevant features from available data attributes. This may require both human resources in terms of intuition and creativity, and an iterative approach to running a learner, analyzing the results and modifying the features and/or the learner (Domingos, 2012). There is no general rule for how to do this which suits every situation or application. There are, however, a number of strategies and tools which can be used as starting points.

A key element of feature engineering in an applied machine learning problem is of course *application domain knowledge*. Existing knowledge should be incorporated in the representation so that a learning algorithm can exploit it (Guyon and Elisseeff, 2003). This becomes increasingly important if the application is very specific and/or if the amount of data available is not huge.

Natural Language Processing

Natural language processing (NLP) is, broadly speaking, computational techniques which make use of language knowledge to process text data. What type of language knowledge to use depends on the application. A simple algorithm may only need to know how to construct words from letters, whereas a more complex algorithm can incorporate a grammar model describing how to correctly put different types of words together in a sentence.

A commonly used representation of natural language is the bag-of-words model. In its most simple form each document, i.e., text data instance, is represented by a feature vector in a vector space with one dimension per word occurring at least once in the entire corpus (a collection of documents). A feature value for a given document can be either binary (1 if the corresponding word is in the document, 0 if it is not), the count of the number of occurrences of the word, or a count score weighted by the length of the document and how common the word is in the whole corpus. This representation does not convey any information regarding the ordering of the words (Manning et al., 2009).

In some NLP applications there is no need to differentiate between different forms of the same word, e.g., *bird* and *birds*, or *talk*, *talking* and *talked*. If only the stem of the word is of interest in the application, any additional affixes such as *-s* or *-ing* are just noise in the data. In order to reduce the amount of such noise, a procedure called *stemming* can be used. Stemming is performed by morphologically parsing each word and identifying the word stem and the affixes (if there are any).

METHODOLOGY

This section describes and justifies this work's methodology of each part of the process. This process can be divided into five phases: 1) domain knowledge acquisition, 2) data acquisition and processing, 3) feature construction, 4) model construction and training, and 5) visualization and evaluation.

Domain Expert Workshop

Due to the lack of written material on the subject of communication efficiency in the air combat domain it is essential to use domain experts as a source of knowledge, in order to understand the problem and gain additional insight on the matter. A workshop was conducted in an effort to acquire such knowledge in a structured manner (Brynielsson et al., 2013).

The participants were four FLSC researchers and three domain experts with both pilot/FC experience (3,000 + 1,000 multirole fighter flying hours, 1,000 commercial aviation flying hours, and 10 years' of FC experience) and knowledge of the FLSC simulation environment (535 flight simulator training facilitation weeks during 16 years). After a brief presentation of the purpose of the workshop, each participant got the time they needed to write down keywords and concepts related to air combat communication efficiency and inefficiency on individual post-it notes. One at a time, the three domain experts explained what they had written and placed the notes on a whiteboard. These explanations sometimes included real world examples and often resulted in follow-up group discussions. Finally all the notes were clustered and each cluster given a name or short description. The results of the workshop, presented in the next section, are of importance for the understanding and motivation of the rest of the methodology.

Data Acquisition and Processing

All of the data used in this work originates from simulations at FLSC. The raw data is primarily of two types, sound recordings and simulator system logs. The sound recordings contain the sounds transmitted on the radio during a simulator session. The system log in question is a record of every instance where a PTT button is pressed or released by a pilot, including timestamps, channel number and which pilot performed the action.

As indicated above, the word content of the radio transmissions is of great interest to the analysis. For this reason the sound recordings need to be transcribed. For simplicity it was decided that the use of automatic speech recognition was to be omitted in this work, as it would potentially introduce too much technical overhead to a study which in many regards is an initial and exploratory one. The task of transcribing the sound recordings was instead performed manually. Apart from extracting the word content, transcribing the recordings also yields a start and an end time for each utterance.

One of the communication aspects highlighted during the workshop was that different types of radio messages have different priority. The ability to automatically label utterances with type based on word content would therefore potentially be useful in a communication efficiency analysis. One obvious approach to this task is supervised machine learning. To make that possible a set of transcribed utterances were manually labeled with type.

The utterance types used were chosen based on comments from the workshop and a later discussion on the matter. These types were *threat calls*, *spike calls*, *declarations*, *picture descriptions*, *tactical calls*, *acknowledgment calls* (utterances which serve only as acknowledgment in response to a previous utterance, containing no additional information) and *other*. The last category is the "none of the above" option and so it contains a variety of utterances, for example certain administrative exchanges such as radio checks. The particular data set used in this study happened to contain only a single instance of the *declaration* type. While the category itself is included in the presentation of the results for completeness, the ability of different models to classify this type of utterance is not further discussed.

The PTT log provides information about who speaks when. The PTT events were merged with the transcribed utterances in order to obtain a set of data instances with complete information regarding who was speaking, what was said and when it was said. Each utterance was thus assigned a speaker. As there is currently no logging of the PTT actions of the FC it was simply assumed that any utterance not assigned to any of the pilots is an utterance of the FC.

The sound recordings with corresponding PTT logs contained a total of 720 utterances. These were the utterances which were assigned utterance type labels. An additional 277 utterances were transcribed from recordings with no PTT logs, meaning they could not be assigned to speakers in a reliable way. The latter set of utterances were therefore only used for limited parts of the analysis.

The air combat language is to a large degree scripted, in a sense somewhere in-between a constructed and a natural language. Hence it is not reasonable to assume that every standard tool of NLP is directly applicable to this language. However, the concept behind a tool may still be relevant and useful. One such commonly used NLP tool is stemming. In this context it would probably be unwise to apply it directly to the text data. For example, there is a difference between *notch* and *notching* where the former is a directive call and the latter a descriptive one. Removing the *-ing* affix and only keeping *notch* would potentially result in loss of important information. However, the idea that some words may look different but still represent the same thing is still relevant. This phenomenon creates noise in the data which one would like to remove. Instead of stemming a domain specific processing tool was designed, from here on referred to as *masking*. The idea of masking is to replace certain words with more generic ones. The words *west*, *east*, *southeast*, etc., all describe a direction. One can be substituted for another and still play the same role in an utterance. From this point of view the difference between *east* and *west* is just noise. By replacing them both with <*direction*> that noise is removed. Other such generic tags used are <*number*>, <*callsign*>, <*location*> and <*brevity*>. The brevity tag is not used for all defined brevity words but only to help "validate" (for lack of a better word) brevity words which are not used very often. Additionally, this tool adds <*start*> and <*end*> tags to the beginning and end respectively of each utterance.

Feature Construction

Due to the scripted nature of the air combat language there are only a few different ways to express a given amount of underlying information for some types of messages. Furthermore, the substantial use of unique brevity words with very specific meanings provides significant clues as to the type of message an utterance may be. Based on these observations a simple bag-of-words model is a reasonable starting point for a language representation. As in this case each text data instance is one single utterance and thus very short, there is little need for the feature model to keep track of the number of occurrences of each word in an utterance. Multiple occurrences of a word in an utterance does not necessarily mean anything in terms of what type of message it is. This potential noise is avoided by defining the feature values as binary.

The bag-of-words model is constructed from a set of training data. In order to limit the dimensionality of the feature space, a word is made a feature only if it occurs at least N times in the training data. When applying the model to unseen data, new words encountered are simply ignored. The point of this design is to make sure all utterances are in the same feature space and a feature space recognized by the trained classifiers. For the construction of this feature model the original utterance data instances were pre-processed through *masking*. However, the <*brevity*> tags were in this particular case omitted, as the brevity words often contain important information regarding the type of utterance. For the minimum word count parameter N=2 was used.

To study the communication on a larger time scale, for example over the course of one simulated scenario, other features are needed. *Channel time usage* is one such feature, which is here defined as the proportion of the time interval $(t–h,\ t+h)$ used for talking, where t denotes a moment in time and h specifies the size of the time window taken into

consideration. *Utterance frequency* is another feature of possible interest. It is similar to channel time usage but only counts the number of utterances within the specified time window. It is essential to consider these two features together as they complement each other concerning which information they convey, due to the fact that the utterances vary significantly in length.

Model Selection and Training

One aspect of communication efficiency is to talk about the right thing at the right time and make sure no important information is drowned in less important one. It is therefore of interest to investigate how, and with what accuracy, utterances can be automatically classified with respect to priority type. One approach to this problem is to construct classification rules by hand, based on knowledge about the air combat domain and language. Two types of such rules were used. The first type of rule defines a set of words which must all be contained in the utterance, and one set of words which cannot be contained in the utterance, in order to count the utterance as a match to the rule. One of the two word sets may be empty. The other type of rule defines only one set of words. All words of an utterance must be in this set of words for the utterance to match the rule. In total 30 rules were used. To some extent these rules were designed to favor precision over recall. The rules were tested one by one in a set order and in case of a match no more rules were tested. As a consequence, the ordering of the rules is also significant and, in a sense, part of the model. If no rule gave a match the utterance was classified as the type *other*.

The alternative to handwritten rules is machine learning. The machine learning models used in this study are based on decision trees. A single tree, taking the entire feature space into account at each node, was trained with the *C4.5* algorithm (Quinlan, 1993). Additionally, the tree based ensemble learning method *random forest* was used. A random forest model is a collection of decision trees which each is trained by considering a randomly sampled subset of the total set of features at every node (Breiman, 2001). In this case the collection consisted of 100 trees, with each tree considering 20 features at each node.

A third option is to combine the two above described models. The idea is to exploit the high precision of a few handwritten rules to filter out utterances which are relatively easy to classify. Utterances not matched by any rule are instead classified by a machine learning model. This was done with 27 out of the original 30 rules. The training of the combined model was done by filtering the training set through the rules and thereby only train the machine learning models on unmatched utterances.

Subject Matter Expert Evaluation

For a human to be able to evaluate the analyzed communication, the features and the types of utterances need to be presented in a comprehensive way. In this work, this is done through visualization. A visual representation can also support the human understanding by, for example, conveying information about the context of the data and/or features in question. In order to investigate the relevance and the pedagogical aspects of the communication analysis, the graphical visualizations were presented during a workshop including one pilot/instructor (3,000 multirole fighter flying hours) as subject matter expert (SME) and two researchers with expertise in communication and pedagogics. The workshop procedure was to first show a specific simulation training sortie to help the SME and the researchers understand the sortie and the outcome of it, and then discuss the graphical presentations of the analyzed data from that specific sortie (see Figures 1–3) through inquiring:

- After what type of training is it relevant to do this communication analysis, and when is it not relevant?
- When and how could a training leader/instructor use the communication analysis?
- Are the graphics sufficient and usable, or are there other ideas on how to present the analysis?

RESULTS

The first part of this section provides a summary of the results of the domain expert workshop, which was conducted as an initial effort to better understand the problem at hand, followed by the results of the feature engineering efforts and the use of machine learning algorithms, with statistical measures and visualizations where applicable.

Domain Expert Workshop

As a result of the initial workshop, three main categories of communicative elements related to efficiency were identified. The first category was labeled *Rules* or *Grammar* (grammar in the sense of how to correctly put together utterances and chains of utterances, not related to the normal grammar of the language spoken). This category includes elements such as proper use of callsigns, response acts, etc. The second category was labeled *Terminology*, with focus on what words to use and how to pronounce them. The third category was labeled *Guidelines* and describes a set of rules which are softer compared to those in the first category. It includes keeping the speech compressed (if something can be said with one word it should not be said with a full sentence), clear and concise. One should speak only if it is necessary, not just because there would otherwise be silence, and only about things that are not already known (for example visible on a display). There is often a set priority order for different types of messages. Threat calls usually have the highest priority, followed by spike calls. Next is declarations and then picture descriptions. Everything else has lower priority.

Priority Type Classification Results

The utterance types are in this paper referred to by numbers according to the following: 1) threat call, 2) spike call, 3) declaration, 4) picture description, 5) tactical call, 6) acknowledgment call, and 7) other. Classification by handwritten rules was evaluated on the whole set of type labeled utterances. The machine learning models were trained using 70% of the type labeled data, and tested using the remaining 30% of the data. The best results for the combined model were obtained when random forest was used as the machine learning component. The combined model was trained and evaluated on the same data sets as the pure machine learning models. Precision and recall for each class and classification model are shown in Table 1.

Table 1. Precision and recall for each class and classification model.

		Class						
		1	2	3	4	5	6	7
Precision	Rules	1.00	0.95	1.00	0.89	0.98	0.99	0.12
	Decision tree	1.00	1.00	0.00	0.85	0.96	0.64	0.67
	Random forest	1.00	1.00	0.00	0.81	0.85	0.70	0.83
	Combination	1.00	1.00	0.00	0.87	0.93	1.00	0.83
Recall	Rules	1.00	1.00	1.00	0.60	0.54	1.00	0.97
	Decision tree	1.00	0.67	0.00	0.94	0.82	0.96	0.50
	Random forest	1.00	0.67	0.00	0.83	0.80	0.96	0.63
	Combination	1.00	1.00	0.00	0.94	0.89	1.00	0.63

Visualizations for Feature Evaluation

The utterance frequency and channel time usage features are designed to represent the communication over the time of one simulator run. Figure 1 shows these features as a time series. Utterance frequency is represented by dots and a dashed line, channel time usage is represented by a solid line. In this case h=15 seconds was used (i.e., the total length of the time window was 30 seconds). Of particular interest in these graphs are the location of peaks and periods of silence. They also provide an initial impression of the amount of variability over the time.

To add some context and aid the evaluation of these features the distribution of utterance types is shown in Figure 2. Each dot represents an utterance. This shows both what types of utterances are used during different time stages of the simulated scenario, and what types of utterances are used in close proximity to each other. The vertical lines represent two pilots attempting to transmit at the same time, which is a potential sign of less-than-optimal communication flow.

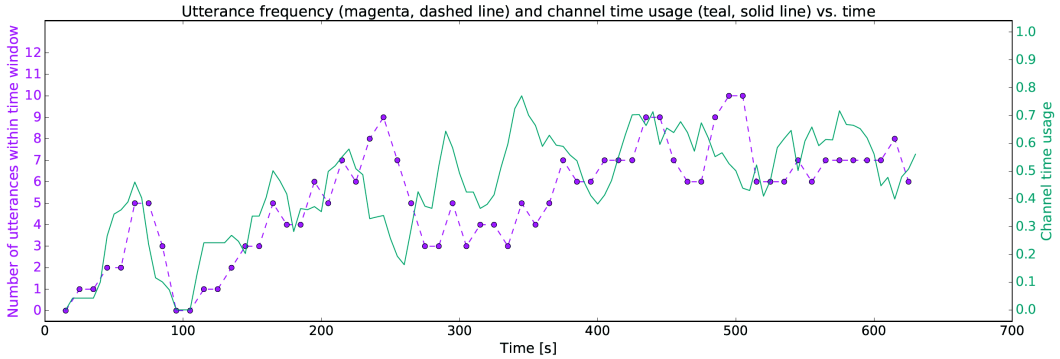


Figure 1. Utterance frequency and channel time usage over time.

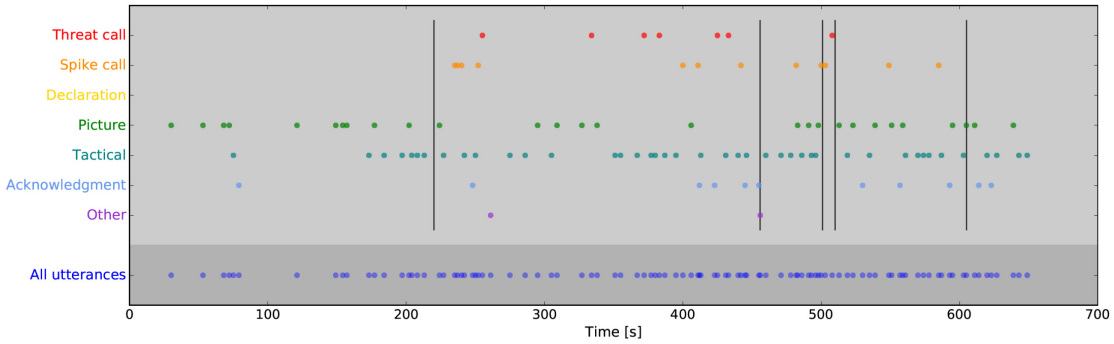
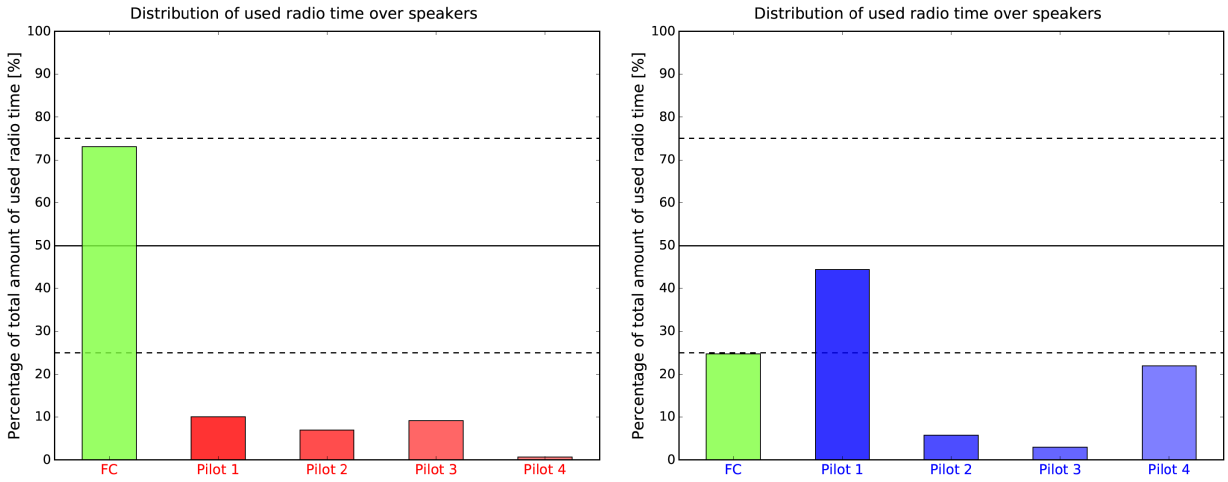


Figure 2. Utterance type distribution over time.

In addition to the presentations above, Figure 3 shows the distribution of used channel time over speakers. Each participant has a role to play in the group, and different roles (here represented by "FC," "Pilot 1," "Pilot 2," "Pilot 3" and "Pilot 4") require different amounts of verbal communication. Of particular interest is in this case to identify if someone appears to have been talking significantly more or less than expected, given their role in the group.



(a) Offensive side. (b) Defensive side.

Figure 3. Distribution of used channel time over speakers.

DISCUSSION

This paper has attempted to answer questions on several levels which will be discussed below in terms of the application and domain centric aspects, and the technical aspects concerning the specifics of machine learning.

Classification

Due to the scripted nature of the air combat language there is reason to believe that classification of utterance types should be a relatively easy task, even if the vocabulary or structure used in reality is not always completely by the book. For certain types of messages the brevity words are so specific that they can be used to formulate classification rules by hand. As indicated by Table 1, the developed rule based classifier performs very well on classes 1, 2, and 6, whilst classification of classes 4, 5, and 7 turns out to be more difficult. This is due to the fact that the language variability within the latter three classes is significantly larger, as is the language overlap between them. For reasons already explained, class 3 is ignored in this discussion.

When it is not obvious how to construct sufficient rules by hand, a natural way forward is machine learning. The two machine learning models tested in this study give a better classification result for classes 4, 5, and 7 but perform worse on classes 2 and 6, compared to the rule model. The performance of the single decision tree is fairly similar to that of the random forest model over all. The former has somewhat better results for classes 4 and 5 but worse for classes 6 and 7.

In an effort to exploit the strength of both the rules and the machine learning a combined model was constructed. The best result was achieved by using random forest as the machine learning component. Taking both precision and recall into account, the combined model performs as well as or better than each other model for every class. This model has a mean precision of 0.93 and a mean recall of 0.91. The idea behind training the machine learning model only on instances not matched by any of the rules is to focus the learning on the concepts which the simple rules cannot deal with. A potential problem with this approach is that the amount of training data visible to the machine learning model is reduced. Further study with more data would be particularly beneficial for evaluation of the combined model.

Temporal Features

In order to evaluate the utterance frequency and channel time usage features they have been visualized as time series over the duration of one simulator run. Looking at such a graph gives a rough idea of the amount of radio traffic over time. A closer study requires more context. A prototype concept for visualization of utterance types is provided as an example. With this combination of information it is possible to both identify trends and interpret them in terms of communication content.

The context dependency is important to keep in mind when comparing graphs from different simulator runs. On a detailed level the graphs may only be comparable if the contexts are comparable. These features may therefore be of particular interest when studying multiple runs of the same simulated scenario, or to compare communication on opposing sides in the same simulator run.

Related to channel time usage is the distribution of the used channel time over speakers. Despite its simplicity, this distribution can give a hint as to who has taken what role in the group. If someone, be it the FC or a pilot, takes a role they are not meant to have, the distribution of radio time may highlight the issue. In the example provided in Figure 3 there is a significant difference between the two sides, which in this case were fighting against each other. The defensive side distribution in particular deviates from the general expectation and would potentially, depending on the tactical context, be an interesting evaluation point.

A Note on Combining Feature Models

Throughout the study features of different types, or related to different aspects of communication efficiency, have been treated separately. One advantage of this approach is that it makes the resulting graphics relatively easy to interpret. This is particularly important in this study as it is not entirely clear what aspects of the radio communication are the most relevant ones to study. It is hoped that by looking at a few closely related features at a time, new insight can be gained.

Combining results to better understand them, as described above, is a small step towards combined models. An extension to this is to let the result of one model be a feature in the next, e.g., a well-trained and reliable classifier can give each utterance a new feature representing the utterance type. Upon new insight and a wider range of features, more complex models can be built.

However, one must not forget that the end goal is for someone to learn something practical about their own or their group's behavior. A non-transparent or very complex analysis, with a result that is difficult to interpret and relate back to reality, is not necessarily useful.

Subject Matter Expert Evaluation

As mentioned in the methodology section, the relevance and the pedagogical aspects of the communication analysis were investigated during a workshop including three SMEs: a pilot/instructor, a communications expert, and a pedagogics expert. The graphical visualizations were presented, and questions regarding relevance, pedagogics, and visualization were posed.

How the used channel time is distributed over speakers will depend on the training scenario, the different pilots' experience levels, and their roles in the group. The graphics showing this distribution (see Figure 3) were deemed to be relevant for any kind of training scenario since it serves as an eye-opener for the individual pilot, for the team, as well as for the instructor. That is, a bar that stands out compared to previous training or the general expectation can stimulate further discussion regarding the particular anomaly and can also be a trigger for discussing the communication in more general terms. The visualizations of utterance frequency, utterance type distribution, and channel time usage over time according to Figures 1–2 were mainly deemed relevant in terms of pinpointing when and how frequently the trainees tried to transmit at the same time.

CONCLUSIONS

The purpose of this study has been to investigate which features of air combat communication are of interest from an efficiency point of view, and what parts of such an efficiency analysis can be performed automatically. It has been shown that the bag-of-words feature model used for classifying utterances serves its purpose well. The classification model combining handwritten rules with the random forest machine learning model performs well, with a mean precision of 0.93 and a mean recall of 0.91.

Channel time usage and utterance frequency are features better suited for assessing communication behavior over time. Further study, with larger amount of data, is needed to find context dependent baselines which can be considered normal or efficient. Significant deviation from such a baseline would potentially indicate inefficiency. Distribution of used radio time over speakers says something about group dynamics and roles with respect to communication. Sometimes a deviation from general expectation will be easily explainable by the tactical context. If it is not, however, the explanation for the anomaly may well be of interest from an efficiency point of view.

Albeit not the final evaluation of its relevance for simulation training, the SME evaluation of the visualization prototypes showed that the concepts have the potential to enhance air combat debriefings. Visualization of the communication is important in order to spark discussions on issues or team performance aspects which would otherwise be likely to fall between the cracks.

Concerning the possibility of automatically performing the efficiency analysis, parts of the process can indeed be automated. Extraction of features such as channel time usage and utterance frequency is one example. Another is the priority type classification of utterances, which this study has shown to be possible with relatively high precision. The graphics presented in this paper require a human to interpret them. It is therefore not a fully automatic process for efficiency analysis. It is, however, a step towards such a process in terms of finding relevant data representations. One must also keep in mind that in this case the whole purpose of analyzing the data is to aid humans in evaluating performance and learning something from it. Completely excluding the human from the process is therefore not necessarily ideal. At what stage would it be optimal to bring in the human is a question for a future study.

REFERENCES

Alpaydin, E. (2010). *Introduction to Machine Learning* (second ed.) Cambridge, Massachusetts / London, United Kingdom: MIT Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brynielsson, J., Nilsson, S., & Rosell, M. (2013). Återkoppling från sociala medier vid insatsledning [Feedback from social media during crisis management]. Technical Report FOI-R--3756--SE, Swedish Defence Research Agency, Stockholm, Sweden.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Lahtinen, T. M. M., Huttunen, K. H., Kuronen, P. O., Sorri, M. J., & Leino, T. K. (2010). Radio speech communication problems reported in a survey of military pilots. *Aviation, Space, and Environmental Medicine*, *81*(12), 1123–1127.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, United Kingdom: Cambridge University Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann Publishers, Inc.

Svensson, J. & Andersson, J. (2006). Speech acts, communication problems, and fighter pilot team performance. *Ergonomics*, *49*(12–13), 1226–1237.

Swedish Armed Forces (2007). Handbok Flygengelska [Flight-English Handbook]. M7748-504022, Stockholm, Sweden.