# Classification of Air Traffic Controller Utterance Transcripts via Warm-Start Non-Negative Matrix Factorization

Marco Enriquez
The MITRE Corporation
Center for Advanced Aviation System Development (CAASD)
McLean, Virginia, USA
menriquez@mitre.org

***Abstract*—Air traffic control voice (i.e., utterance) transcript data is often underutilized in the context of airspace analysis, despite its increasing availability. This provides an opportunity, as controller utterances contain operational directives which typically have to be inferred from aircraft trajectory state data. Direct knowledge of such directives would enable, enhance or expedite various analyses efforts such as Area Navigation (RNAV) procedure assessment, airspace redesign or air traffic controller workload studies. Despite the fact that transcribed utterances can be free-form (or loosely structured), controller utterances fit into a finite number of categories. To this end, we propose using domain knowledge to create an effective Warm-Start strategy for the Non-Negative Matrix Factorization (NMF), which in turn can be used to automatically categorize utterance transcripts. Exploiting domain knowledge yields two key insights: the number of topics to expect from the collection of utterances, and the partial list of terms to expect from the aforementioned topics. Knowing the number of topics implies the desired rank of the factorization and hence removes a significant free parameter from the NMF algorithm. Coupling the partial list of topic terms can be used to create the (warm-start) initial guess matrix to the NMF algorithm. Using human and machine transcribed voice data, we show that our approach closely matches manually labeled (i.e., by subject matter experts) utterances, and compares favorably against the Partially Labeled Latent Dirichlet Allocation (PL-LDA) algorithm. Furthermore, we “stitch” labeled utterances from the DAWGS Standard Instrument Departure (SID) and FLCON Standard Terminal Arrival Route (STAR) to their corresponding trajectories, and propose a novel procedure analysis methodology via geospatial utterance heatmaps.**

## I. Introduction

Voice (“utterance”) data is an integral part of modern aviation, as pilots rely on verbal guidance from air traffic controllers during all phases of flight. Acquisition of utterance data has been more common world-wide, as various research institutions form alliances with local airports and air traffic control centers (e.g., [5], [6], [16], [25]).

With the increasing availability of utterance data comes the tremendous opportunity to mine and extract content from such data. Typically, aviation analysts only have access to trajectory states, from which they must deduce directives such as a speed adjustment or a vector. Though algorithms can be written to identify such directives from trajectory data, ensuring accuracy of such deductions can be difficult. Having access to utterance data would significantly lessen ambiguity from such assertions. However, since each of the 22 air traffic control centers in the National Airspace System (NAS) can generate hundreds of thousands of utterances in one day, manual analysis of utterances is undesirable.

In this paper, we consider the Non-negative Matrix Factorization (NMF) algorithm to automatically classify and label utterance transcripts given user-defined topics, such as “speed”, “altitude” or “weather”. The NMF is an optimization-induced matrix decomposition technique that decomposes a non-negative matrix into the product of two non-negative matrices. In the literature, the NMF has been used for identifying “topics” from a corpus, as well as roughly identifying what percentage each topic appears in a specific document belonging to the corpus. We note that the term “topic” in the document analysis literature refer to a collection of words or terms that pertain to one subject. For example, the “altitude directives” topic might contain the terms “descend”, “climb” and “flight level”. Having access to the utterance topics, in turn, supports multiple air traffic management analysis activities such as: providing inputs to air traffic controller workload models and understanding where flights get vectored off Optimized Profile Descents (OPDs).

Ensuring classification accuracy with the NMF can be challenging, however. Since the NMF is an explicit optimization process, the supplied initial guess strongly affects convergence to a solution. Furthermore, the rank of the factorization matrices is typically a required input to the NMF algorithm, and can play a significant role in the quality of the solution. Our main contribution in this paper is to propose a “warm-start” strategy to combat the two aforementioned challenges. In the optimization literature, “warm-start” strategy is a methodology that supplies an optimization algorithm with a superior initial guess. The warm-start strategy we develop leverages domain knowledge to both create a reasonable hypothesis for the rank of the NMF factorization matrices and create a strong initial guess to the NMF algorithm.

Our proposed algorithmic approach assumes that two main challenges have been addressed. First, since utterance data comes in one continuous stream, it must be segmented into utterances to facilitate transcription. Though we will not focus on acoustic segmentation in this paper, we note that many works have proposed various algorithms to tackle such problems (e.g., [7], [9] and [10]). The second challenge involves transcribing the segments of voice data to text files.

Both research and commercial voice transcription software exist, and can do an adequate job of transcription given that a certain level of fidelity is met and that the speaker properly enunciates words. Since controller utterance data gathered from facilities may not always satisfy the two aforementioned assumptions, the software-generated transcription will likely contain errors. Granted that these errors are not severe, we can leverage the NMF to analyze the transcriptions. Before proceeding, we note that some of these challenges regarding transcription and classification will be addressed once the Federal Aviation Administration (FAA) fully implements their Data Communications (DataComm) initiative, which would enable pilots and air traffic controllers to replace routine and repetitive communications with digital textual messages [1], [19]. To our knowledge, however, the FAA still plans for air traffic controllers to use voice messages for non-commonplace communications.

Though many other prominent topic identification algorithms exist in the literature (e.g., Latent Semantic Allocation [8] and Latent Dirichlet Allocation [4]), we choose the NMF for the following reasons. First: we claim that the algorithmic and mathematical framework of the NMF easily supports "warm-start" strategies, allowing analysts to exploit aviation domain knowledge to improve transcript classification quality. Second: the NMF is readily interpretable due to the non-negative weights it produces, which is useful for classifying the utterance transcripts. (We show concrete examples of utterance classification via the NMF in the third section of this paper.) Third: we show that using a simple variant of the NMF, when coupled with our proposed warm-start strategy, is enough to obtain great classification results. Specifically, in this work, we use Lee and Seung's NMF algorithm [17] which only relies on matrix-matrix multiplications[1]. It is worth noting that the aforementioned algorithm is also *minimally* reliant on tuning parameters. To the author's knowledge, these desirable features are not all present in any other topic detection algorithm. Indeed, Latent Semantic Allocation (LSA) requires a singular value decomposition (SVD). It is well known that the SVD incurs an $O(n^3)$ computational cost in general; for large datasets, we must rely on less expensive algorithms that can approximate the singular vectors and values of a matrix (e.g., [12]). Further, due to the orthogonality condition of the singular vectors, LSA cannot account for multiple meanings of a word (i.e., polysemy). Latent Dirichlet Allocation (LDA), on the other hand, does not offer an obvious way to incorporate supervised terms into its learning process. LDA may also require hyper-parameter tuning to improve classification results. Hyper-parameter tuning aside, there are variants of LDA that incorporate supervision; to our knowledge, the LDA variant most similar to the approach we propose here is the "Partially Labeled LDA" (PL-LDA) [27]. In addition to the hyper-parameters from the LDA algorithm, the PL-LDA introduces four extra algorithmic parameters: the number of background topics, the number of topics per label, a term smoothing term and a topic smoothing term. We will show that the "warm-start" NMF (WS-NMF) performs favorably compared to the PL-LDA.

The paper is organized as follows. The second section discusses the methodology of this work. The third section analyzes human and machine transcribed utterances, and presents accuracy results. The fourth section associates each utterance label to trajectory data, and proposes various areas that could be researched as a result. The final section concludes.

## II. Methodology

Non-Negative Matrix Factorization can be attributed to Lawton and Sylvestre [15] who decomposed a continuous curve into the non-negative linear combination of two functions. Lee and Seung [17] popularized the NMF by proving that straightforward multiplicative update rules suffice to compute the NMF. Since Lee and Seung's work, more complex variants of the NMF algorithm appeared in the literature, such as sparseness-constrained NMF [13], NMF via alternating non-negativity constrained least squares [14], and NMF via projected gradient methods [20]. Though our warm-start strategy is applicable to any NMF algorithm, we will focus on Lee and Seung's algorithm for the remainder of this paper.

Mathematically, the NMF problem can be stated as follows: given a matrix $A \in \mathbb{R}^{m\times n}_{+\cup\{0\}}$ (note that the subscript implies $A$ only contains non-negative entries) and the desired rank $r$, find matrices $W \in \mathbb{R}^{m\times r}_{+\cup\{0\}}$ and $H \in \mathbb{R}^{n\times r}_{+\cup\{0\}}$ such that $A \approx WH^T$. NMF has been used for many applications such as facial feature decomposition (e.g., [11] and [26]) and document topic identification (e.g., [17] and [31]). We rely on the latter in order to classify utterance transcripts. In the context of document topic identification, each column of $A$ (referred to in this setting as the *document term matrix*) corresponds to a word or term found across all documents, while each row of $A$ corresponds to a specific document. The values on the $i^{th}$ row of $A$ then reflect how often specific terms appears in document $i$. Performing the NMF will then induce a representation of each document (i.e., each row of $A$) as a linear combination of $r$ topics, with the columns of $W$ illuminating the $r$ identified topics and with the rows of $H^T$ containing the weights on each topic.

Typically, we compute the NMF of a matrix by solving the following optimization problem for a fixed positive integer $r$:

$$\min_{W\in\mathbb{R}^{m\times r}_{+\cup\{0\}},H\in\mathbb{R}^{r\times m}_{+\cup\{0\}}} \|A - WH^T\|_F^2 \tag{1}$$

where the subscript $F$ denotes the Frobenius norm, defined as the following for a matrix $X \in \mathbb{R}^{m\times n}$:

$$\|X\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{i,j}{}^2}.$$

To solve the optimization problem (1), Lee and Seung [17] proposed the following alternating multiplicative update rules:

$$H \leftarrow H\frac{(W^TA)}{(W^TWH)} \qquad W \leftarrow W\frac{(AH^T)}{(WHH^T)}. \tag{2}$$

[1] We note that the computational complexity of matrix-matrix multiplications is $O(n^k)$ for some $k$ between 2 and 3 (e.g., Strassen's algorithm, with computational complexity $O(n^{log_2 7})$). Furthermore, automatic software optimization packages exist to further accelerate computations, such as ATLAS [30]. Finally, we remark that matrix-matrix multiplication trivially scales as it is an embarrassingly parallel computation.

**Algorithm 1** Warm-Start NMF (WS-NMF) for analyzing utterance transcripts.

```
1:  **Input :** $r$ groups of terms, each group representing a topic
2:  **Input :**$A \in \mathbb{R}^{m \times n}_{+\cup\{0\}}$, with $m$ terms across $n$ transcripts
3:  **Input :** weight and pertinence parameters $\alpha, \delta \in \mathbb{R}_+$
4:  Initialize $W^0 = zeros(m, r), \quad H^0 = rand(r, n)$
5:  **for** $j = 1 ... r$
6:      **for** $i = 1 ... m$
7:          Set $W^0_{i,j} = \alpha$ if term $i$ is in topic $j$
8:  $\bar{W}, \bar{H}$ = **NMF**($W^0, H^0$, $max\_iter$)
9:  **for** $k = 1 ... n$
10:     Label transcript $k$ with topics where $\bar{H}_{k,:} > \delta$
```

**Algorithm 2** Lee and Seung's Multiplicative Update Rules for computing the NMF.

```
1:  **Input :** $W^0$, initial guess for topic matrix
2:  **Input :** $H^0$, initial guess for weight matrix
3:  $W = W^0, \quad H = H^0$
4:  **while** (not converged) **or** $iter < max\_iter$
5:      $H \leftarrow H \frac{(W^T A)}{(W^T W H)}$
6:      $W \leftarrow W \frac{(A H^T)}{(W H H^T)}$
7:   **Output :** $W$, $H$
```

(Note that the matrix-matrix divisions are conducted element-wise in the equations above.) Compared to other NMF variants mentioned above, Lee and Seung's algorithm is desirable because it is simple to implement, as it only requires matrix-matrix multiplications. Algorithms capable of matrix-matrix multiplication have been implemented for various platforms – using both the serial and parallel computing paradigms – and are readily available (e.g., BLAS [3], LAPACK [2], and NumPy [24]). Further, Lee and Seung's algorithm does not rely on tunable parameters and is provably convergent to a local minimizer. We state Lee and Seung's main result here:

**Theorem 1.** *The Euclidean distance $\|V - WH\|$ is non-increasing under the update rules (2). The Euclidean distance is invariant under these updates if and only if $W$ and $H$ are at a stationary point of the distance.*

For a proof of the above theorem, as well as alternate NMF objective functions which can also be minimized via multiplicative update rules (e.g., the Kullback-Liebler divergence), we refer the readers to [18].

It is well known that the NMF suffers from two main drawbacks: first, the rank $r$ of the factorization matrices (i.e., $W$ and $H$) need to be known *a-priori*. It is clear from the NMF problem that the chosen rank $r$ plays a significant role in the quality of the induced factorization. Second, unlike famous matrix decompositions such as the Singular Value Decomposition (SVD), NMF does not yield a unique decomposition. Since the NMF is an optimization process, the starting guess provided to the optimizer greatly affects the algorithmic solution.

We claim that we can bypass the two above challenges, however, due to the nature of the problem we wish to solve and our domain knowledge. Since we are trying to classify utterance transcripts into a few *known* categories, the value of $r$ should also be known to us. Further, domain knowledge also allows us to generate words that partially constitute each topic bin. We can translate this into a superior initial guess for the topic matrix $W$. Algorithm 1 details the generation of the warm-start initial guess to the NMF algorithm, as well as how to use the output of the NMF algorithm to classify transcripts.

In algorithm 1, we begin by supplying four inputs. First, we need to identify the pertinent topics we wish to classify the utterances with, and supply a few related terms per each topic. Second, we need to generate the document term matrix from the utterance transcripts. For the analysis here, we consider a "bag-of-words" model (i.e., we assume that ordering of words in the transcripts do not matter). We count the occurrence of each unique term across all transcripts, excluding commonly used words that add no contextual value. These term counts can then be transformed via schemes that illuminate specific terms' importance to the corpus, or hidden inter-term relationships. Furthermore, we create "n-grams" to consider multiple word terms in the analysis. (For example, a 2-gram segmentation of the sentence "The sky is blue" will yield the terms "the sky", "sky is" and "is blue".) We use the term frequency–inverse document frequency (tf-idf) update [29] for this work, whose output is stored in the document term matrix in a fashion described earlier. Third, the parameter $\alpha$ represents the weight we wish to give the entries in the initial topic matrix, $W^0$. Fourth, the parameter $\delta$ quantifies the minimum value that the converged weight matrix's entry (i.e., $\bar{H}_{i,j}$) must have in order for it to be considered significant. Since the value chosen for $\alpha$ affects the magnitude of the matrices $\bar{W}$and $\bar{H}$, the appropriate value for $\delta$ should be chosen for each $\alpha$. (We note that these parameters are not hard to select in practice; for all the tests we consider here, we set $\alpha = 1$ and $\delta$ to be machine-epsilon, $\epsilon_{mach}$.) Given these inputs, lines 4 to 7 then describe how to initialize the topic matrix $W$, given a guess for the topic bins. In turn, the initial guesses $W^0$ and $H^0$ are passed to the NMF algorithm in line 8. After the NMF call, we take the topic weight matrix $H$ and perform the classification; indices where values of the $k^{th}$ row of $\bar{H}$ is greater than some tolerance $\delta$ correspond to the topics utterance $k$ is associated with.

The NMF algorithm we use in algorithm 1 can be any algorithmic variant capable of explicitly receiving an initial guess for the weight and topic matrices. Algorithm 2 implements Lee and Seung's NMF algorithm to minimize objective function 1. In algorithm 2, the update schemes (2) are implemented in lines 5 to 6. We repeat the update rules until the difference between $A$ and $WH^T$ are small in norm, or a certain number of iterations have been reached.

Though not mentioned in algorithm 1, the topic matrix $\bar{W}$ (i.e., the topic matrix after convergence of the NMF algorithm) warrants closer examination. Recall that we only supplied the NMF algorithm a *partial* list of topics. Analyzing $\bar{W}$ will yield *extra* terms that belong to each topic, which itself can highlight interesting information. Recalling the example mentioned in the introduction, suppose we initialized the

| Altitude | Speed | Hold | Vector | Transfer |
|---|---|---|---|---|
| climb | mach | hold | vector | contact |
| descend | speed | clearance | resume | |
| cross | knots | | heading | |
| now to | increase | | turn | |
| | reduce | | degrees | |
| | resume | | direct | |
| | | | cleared | |
| | | | proceed | |

Table I: Initial topic list for tests A,B and C.

| Altitude | Speed | Hold | Vector | Transfer |
|---|---|---|---|---|
| maintain | maintain | | dirty | Atlanta |
| flight | able | | | center |
| level | spacing | | | |

Table II: Extra terms identified by both the WS-NMF and the PL-LDA algorithms from Test C. Note that both algorithms identified the exact same set of terms independently.

"altitude directives" topic with the terms "descend", "climb" and "flight level". The appropriate column of $\bar{W}$ might also indicate that the term "maintain altitude" belongs to the "altitude directives" topic.

## III. Utterance Transcript Analysis

In this section, we test our proposed approach by running algorithm 1 on two types of transcribed utterances. Tests A, B and C use algorithm 1 on human-transcribed controller utterances. The data used for tests A, B and C are utterances transcribed from the Atlanta Air Route Control Center (ZTL), taken on April 4, 2012 from 1100Z-0100Z. Across these three tests, we consider audio transcripts from varying flight levels (FL) of both the DAWGS Standard Instrument Departure (SID) and FLCON Standard Terminal Arrival (STAR). We classify transcripts from tests A, B and C with the topics: altitude change, speed change, holding, vectoring or transfer of control. Subject matter experts provided an initial guess of terms which comprise these topics, which are listed in table I. In Test D, we test our algorithm on transcripts generated by Automated Speech Recognition (ASR) software. The voice data from Test D come from a Human-In-The-Loop (HITL) simulation for the Northern California Terminal Radar Approach Control (TRACON). The software Loquendo [21] was used to transcribe the HITL utterances via a statistical language model. We note that this is a more difficult (and realistic) test of utterance classification, as ASR software may mis-transcribe, or drop, certain words during the transcription process. Test D groups the transcripts according to the following topics: altitude change, speed change, heading, direct, frequency and approach. The initial guess of terms which comprise these topics can be found in table III. Further, in test D, we test each algorithm's ability to determine the one dominant topic. Hence, no false-positive metrics were tracked for this test.

| Altitude | Speed | Heading | Direct | Freq. | Approach |
|---|---|---|---|---|---|
| climb | mach | heading | direct | contact | via |
| descend | speed | | cleared | | descend |
| maintain | knots | | | | arrival |
| flight level | increase | | | | runway |
| | reduce | | | | |

Table III: Initial topic list for test D.

Accuracy results, as well as a detailed description of each test, are shown in tables IV and V. Accuracy of the algorithm is determined by comparing the algorithm's labeling of the transcripts to subject matter experts' consensus. The initial guess topic list used is listed in II. The parameters $\alpha$ and $\delta$ were chosen as 1 and $\epsilon_{mach}$ (i.e., machine-epsilon), respectively, for the WS-NMF algorithm. The results from the WS-NMF algorithm can be seen in table IV. Further, we explore specific WS-NMF mis-classifications for test D in table VI. We use the Stanford Topic Modeling Toolbox [28] to perform the PL-LDA for all tests shown here. Preparing data for use with the PL-LDA algorithm was a fairly straight forward process that parallels steps 4 to 7 in algorithm 1; if an initial guess term (from table I for tests A, B or C, or table III for test D) matches an utterance, that topic associated with that term was assigned to the utterance. Otherwise, no topics were assigned to the utterance. Due to the non-sparseness of the topic weights generated by the PL-LDA inference process, the thresholding parameter $\delta$ plays an important role in the classification results, and the false-positive rate. Table V shows the PL-LDA accuracy results for multiple values of the thresholding parameter. We see that the WS-NMF compares favorably to the PL-LDA across all tests, even after using the best thresholding parameter for the PL-LDA algorithm. Also, note that for the human-transcribed utterances, the number of false positives generated by the WS-NMF is considerably less than the PL-LDA. Altering some parameters in the PL-LDA might yield better results; we noticed, for example, that setting the number of allowed background topics for the PL-LDA to 3 increased the accuracy of tests A, C and D by $1-2\%$, decreased the accuracy of test B by roughly $1\%$, and reduced the false-positive rates across all tests by roughly $10\%$. Further increase of the allowed background topics yielded no further improvements, however. The same could be said for the number of topics per label, which was set to 1. Further, for all the examples we consider here, the term smoothing and topic smoothing parameters were both set to $0.01$ – the values suggested by the TMT toolbox examples.

We also note that our proposed approach has similar functionality to the PL-LDA. First, we refer to table VII, which highlights the topic weights given by the WS-NMF and the PL-LDA for select transcripts from test C. We note that both algorithms are capable of identifying multiple topics, as seen in the fourth row of table VII. The WS-NMF, however, produces sparse weight vectors – facilitating classification of the utterances. In turn, this reduced the number of false-positives incurred during classification. We also point out that both the WS-NMF and PL-LDA can easily identify "outlier" transcripts, potentially highlighting non-common communications between the air traffic controller and the pilot. In turn, this can be used for aircraft safety and risk analysis. The WS-NMF algorithm identifies outliers by assigning a zero weight to all topics, whereas the PL-LDA assigns a high weight to the "latent" topic. In table VII, we see how the WS-NMF and

| Test | Description | $n$ | $m$ | WS-NMF Accuracy |
|---|---|---|---|---|
| A | Human-transcribed utterances from ZTL Sector 16. Covers the en-route portion of the DAWGS SID from FL150 to FL230. | 804 | 6735 | 99.20% (fp: 11) |
| B | Human-transcribed utterances from ZTL Sector 32. Covers the en-route portion of the DAWGS SID from FL240 to FL290 | 633 | 5350 | 98.66% (fp: 4) |
| C | Human-transcribed utterances from ZTL Sector 50. Covers the en-route portion of the FLCON star from FL240 to FL340. | 1821 | 11222 | 99.72% (fp: 2) |
| D | ASR-transcribed utterances from a northern California HITL. The software Loquendo was used to transcribe the utterances. | 276 | 1384 | 86.58% |

Table IV: Test descriptions, number of utterances ($n$), number of unique terms in the corpus ($m$) and accuracy rates for our proposed approach, the warm-start NMF (WS-NMF). Note "fp" refers to the number of false-positives identified by the WS-NMF Algorithm. A false positive is counted when an algorithm assigns an incorrect label to an utterance. Since each utterance may contain multiple labels, it is possible for the false positives to outnumber the actual amount of utterances.

| | PL-LDA Accuracy | | | | | |
|---|---|---|---|---|---|---|
| Test | $\delta = \epsilon_{mach}$ | $\delta = 1e^{-4}$ | $\delta = 1e^{-2}$ | $\delta = 1e^{-1}$ | $\delta = 5e^{-1}$ | $argmax(H_{\cdot,:})$ |
| A | 87.80% (fp: 3413) | 87.44% (fp: 3044) | 89.01% (fp: 276) | **90.34% (fp: 203)** | 87.68% (fp: 102) | - |
| B | 86.73% (fp: 2616) | 86.73% (fp: 1910) | 90.68% (fp: 227) | 93.52% (fp: 127) | **94.92% (fp: 32)** | - |
| C | 80.99% (fp: 7630) | 80.89% (fp: 6060) | 90.94% (fp: 902) | **93.85% (fp: 544)** | 84.57% (fp: 281) | - |
| D | - | - | - | - | - | **75.71%** |

Table V: Partially labeled LDA (PL-LDA) accuracy, tests A-D, for various tolerances $\delta$. Best results, for each test, are shown in bold font. Note that, for test D, we only consider an algorithm's ability to identify the dominant topic. This was accomplished by taking the maximum value of the matrix $H$, per row. Note that this methodology was not applied to tests A-C, as those tests try to identify multiple topics, if they exist.

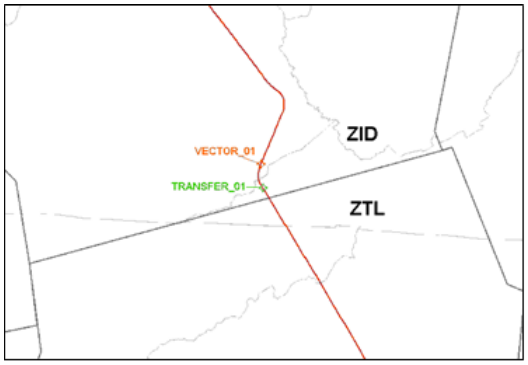


Figure 1: Labeled utterances being affixed to an aircraft trajectory (red line).

the PL-LDA identifies a transcript that belongs to a different topic (namely, weather). Second, we refer to table II, which demonstrates that both the WS-NMF and PL-LDA can learn extra pertinent terms that comprise a topic.

## IV. Affixing Utterance Labels to Trajectories

Once the labels for each transcript are identified, we can "affix" them to the trajectories they are associated with, as illustrated in figure 1. This requires the aircraft ID and the time-stamps associated with each transcribed utterance. Such associations will enable and enhance various analysis efforts, such as an in-depth examination of procedure utilization or analyzing air traffic controller workload. In this section we consider the former, by analyzing trajectories on the en-route portions of the DAWGS SID and FLCON STAR in the Hartsfield-Jackson Atlanta International airport (ATL) airspace on April 12, 2012, at various flight levels.

We begin this analysis by first classifying utterance transcripts according to the topics listed in table I, via Algorithm 1. A simple count of each utterance label, per procedure and flight level range, can be found in Figure 2. We see

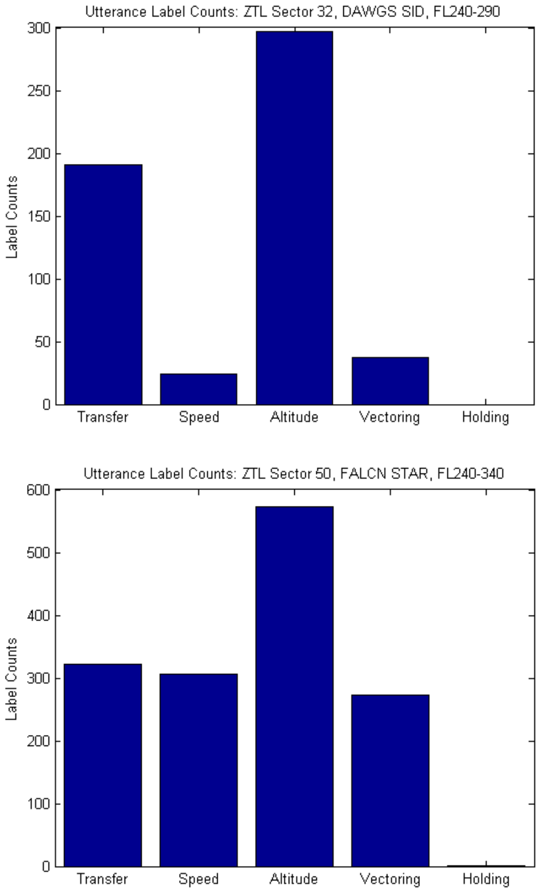


Figure 2: Bar graph showing utterance transcript label counts for two procedures in Atlanta Air Route Control Center (ZTL) on April 12, 2012. The top figure corresponds to the en-route section of the DAWGS SID in sector 32 of ZTL, while the bottom figure corresponds to the en-route section of the FLCON STAR in sector 50 of ZTL.

| | **WS-NMF Topic Weights** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Utterance Transcript** | Altitude | Speed | Heading | Direct | Freq. | Approach | **SME Classification** |
| “two exceed two thirty one turn center roger descend via the center arrival” | 0.6302 | 0 | 0 | 0 | 0 | 0.3698 | Approach |
| “thirty twenty four fifty five descend-and-maintain flight-level one niner zero maintain two six zero knots” | 0.6969 | 0.2874 | 0 | 0 | 0 | 0 | Speed |
| “five ninety one mike nor-cal approach route greater contact maintain five one niner zero” | 0.0510 | 0 | 0 | 0 | 0.9490 | 0 | Altitude |

Table VI: Specific examples of WS-NMF mis-classifications, from test D. Recall that for this test, we wish to identify the dominant topic of the utterance. Though the WS-NMF algorithm assigns a larger weight to a false-positive topic, it also assigns a positive weight to the correct topic. Such mis-classifications may be averted by adaptively selecting $\alpha$ in algorithm 2; this is expounded in the concluding remarks as a topic for future research.

| | **WS-NMF Topic Weights** | | | | | **PL-LDA Topic Weights** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Utterance Transcript** | Altitude | Speed | Hold | Vector | Transfer | Altitude | Speed | Hold | Vector | Transfer | [Latent] |
| “delta 2032 contact atlanta center 121.35 good day” | 0 | 0 | 0 | 0 | 1 | 0.0006 | 0.0027 | 0.0069 | 0.0062 | 0.8387 | 0.1448 |
| “citrus 597 atlanta center cleared direct DIRTY” | 0 | 0 | 0 | 1 | 0 | 0.0003 | 0.0003 | 0.0005 | 0.4854 | 0.0101 | 0.5035 |
| “delta 2143 descend and maintain flight level 240” | 1 | 0 | 0 | 0 | 0 | 0.8831 | 0.0006 | 0.0401 | 0.0003 | 0.0034 | 0.0724 |
| “citrus 204 cleared direct DIRTY descend and maintain flight level 310” | 0.3077 | 0 | 0 | 0.6923 | 0 | 0.6619 | 0.0127 | 0.0004 | 0.3237 | 0.0002 | 0.0011 |
| “ASQ 4688 radar detects multiple scattered cells of moderate to heavy precipitation along the FLCON 7 arrival if you need to deviate advise” | 0 | 0 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0008 | 0.0006 | 0.0 | 0.9985 |

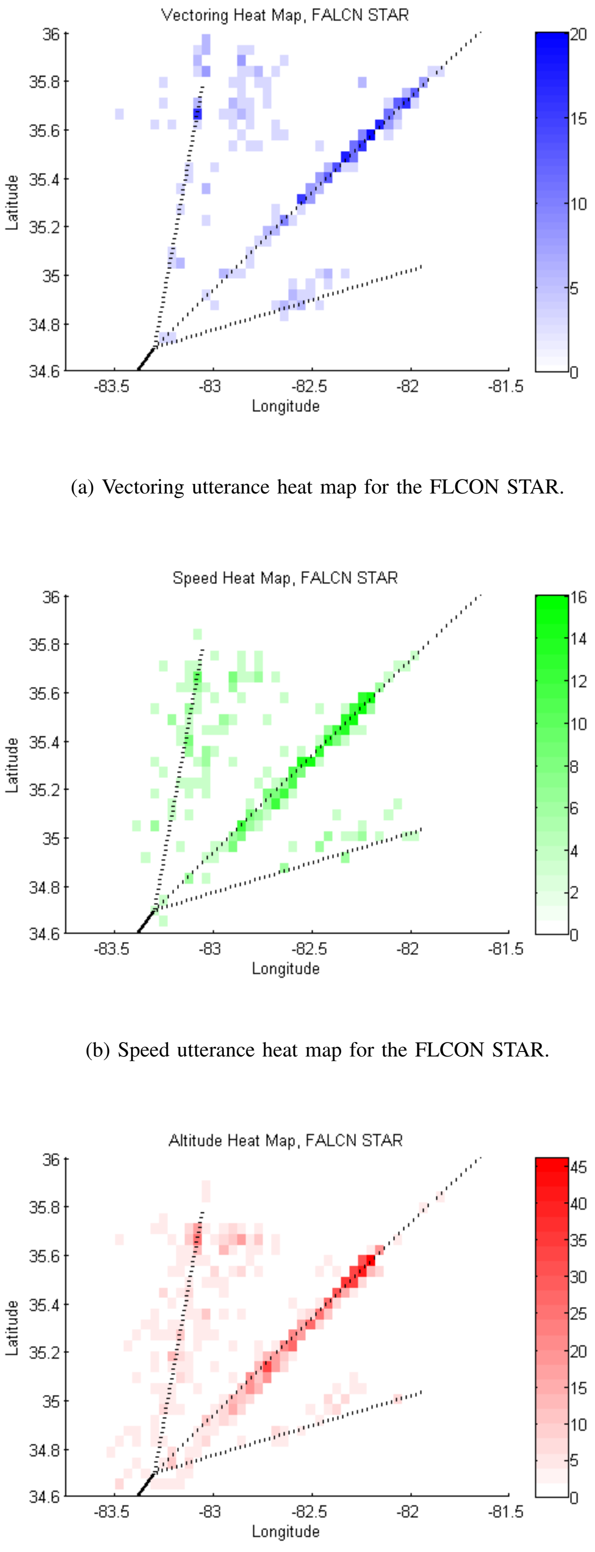
Table VII: Topic weights for selected transcripts from Test C, from both the NMF approach and the partially labeled LDA (PL-LDA). The NMF weights were normalized to sum to 1, per row, to enable direct comparison with the PL-LDA. Further, note that the last column under the PL-LDA section refers to the weight of the latent (background) topic.

immediately that, on April 12, 2012, altitude-related utterances are prevalent in both the DAWGS SID and the FLCON STAR (for the flight levels we consider), though the FLCON STAR has a significant amount of speed and vectoring related utterances as well. In turn, this information can be used to provide insight into the operating status of a procedure. The utterance topic counts can also aid controller workload studies; it has been asserted in [22], and many other studies, that the number of communications and instructional clearances are correlated to higher subjective controller workload ratings. Classified utterances can also be used as inputs to controller workload models.

We can delve deeper into the FLCON STAR utterance data by examining the geospatial bins associated with each utterance – in essence performing a geospatial heatmap analysis. We affix the speed, altitude and vectoring topics to their respective trajectories, and feed the extracted latitudes and longitudes to a geospatial binning algorithm called “geohashing” [23]. Essentially, geohashing creates a two-dimensional histogram of the earth’s surface, where each bin has a standardized boundary and size per precision level. Each geospatial bin has an approximate dimension of $4km \times 5km$, corresponding to precision level $5$ in the geohashing algorithm. Figure 3 highlights the geospatial bins, with the appropriate intensity, overlayed on the FLCON STAR for the speed, altitude and vectoring topics. This information is particularly important, as this will allow us to find potential inefficiencies in new or redesigned procedures. An example use-case might be to study where flights get vectored off optimized profile descents (OPDs). Another example use-case could examine where “outlier” utterances (i.e., utterances which did not map to any of the topic lists) occur.

## V. Conclusion

The increasing availability of pilot-controller communication data, whether in the form of voice data or digital text messages, necessitate technologies which facilitate their analysis. Rapid analysis of such communication data can greatly aid and expedite aircraft safety analysis, procedure analysis and understanding/modeling air traffic controller workload. In this paper, we presented a methodology for classifying transcribed air traffic controller utterances (either manually



(a) Vectoring utterance heat map for the FLCON STAR.

(b) Speed utterance heat map for the FLCON STAR.

(c) Altitude utterance heat map for the FLCON STAR.

Figure 3: Utterance geospatial heatmap analysis on Atlanta Air Route Control Center (ZTL) Sector 50 for the FLCON STAR (black line), FL240-340, on April 12, 2012. Latitudes and longitudes were inferred by affixing utterance labels to trajectories. The obtained coordinates are then grouped into geospatial bins via a geohashing algorithm. The color bars on each figure depict how the color intensities map to the number of utterances.

or via automatic speech recognition software), via the non-negative matrix factorization (NMF). We claim that the NMF is an effective tool for classifying air controller utterance transcripts for three reasons. First, the NMF optimization process lends itself to “warm-start” strategies, allowing us to leverage our aviation domain knowledge. Second, the non-negativity of the topic weights that stem from the NMF are readily interpretable, which is useful for the labeling (classification) process. Third, when coupled with the warm-start approach, a simple variant of the NMF – namely, Lee and Seung’s algorithm – produces excellent classification results. We consider Lee and Seung’s multiplicative update rules for the following reasons: they are easy to implement and parallelize (due to the availability of scientific software capable of matrix-matrix multiplications), they are minimally reliant on tunable parameters, and together, they form a provably convergent scheme. Results on four transcript datasets (from Atlanta Air Route Control Center and a human-in-the-loop simulations) show the promise of our approach; manually transcribed utterances were classified with an $98\%$ accuracy rate and ASR transcribed utterances were classified with an $86\%$ accuracy rate.

We also compared the NMF to the PL-LDA in this work. We show that for static data classification, the NMF has the same inference capabilities as the PL-LDA (such as outlier detection and multiple-topic association). Further, we note that in the tests we consider here, the NMF performed considerably better than the PL-LDA at utterance classification. The NMF also generated significantly less false-positive classifications of utterances. We acknowledge that LDA and its variants (including the PL-LDA) are more suited for data-streaming analysis; after analyzing a training set, the LDA algorithm can more easily approximate the topic distributions of new documents. With the NMF, introducing new documents for classification would likely necessitate solution of another optimization problem.

This work also proposed utterance transcription analysis methodologies that can be enabled or expedited via technologies presented in this paper. After applying the classification algorithm, we can easily monitor utterance topic counts associated with a procedure. This, in turn, could be used in a variety of ways such as: inputs to a controller workload model, tracking the operating conditions of a center, etc. We also showed that affixing the utterance topics to their respective locations and trajectories can enable geospatial analysis of procedures. In the future, we wish to explore adaptive schemes for computing $\alpha$ in algorithm 1. We hypothesize that it may improve classification performance to choose a different value for $\alpha$ for each pertinent entry in the initial topic matrix $W^0$, corresponding to the probability that a term belongs to a topic group. For example, we might consider making $\alpha$ a function of the ASR per-word confidence scores. Relating the probabilistic interpretation of the ASR per-word confidence score to the probabilistic interpretation of the NMF initial guess remains to be completed, and is the topic of future research.

## Acknowledgment

The author would like to thank Genki Kondo, Travis Gonzalez, Rob Tarakan, Ron Chong, Quian Hu, Shuo Chen, Mahesh Balakrishna, Mike Mills, Tom Becher, Eric Hughes and Lashon Booker from MITRE for providing datasets for this work and for providing helpful feedback. This work was fully funded by MITRE's Early Career Research Fellowship.

## References

[1] Federal Aviation Administration. www.faa.gov/about/office_org/ headquarters_offices/ato/service_units/techops/atc_comms_services/ datacomm.

[2] E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof, and D. Sorensen. Lapack: A portable linear algebra library for high-performance computers. In *Proceedings of the 1990 ACM/IEEE Conference on Supercomputing*, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[3] L. S. Blackford, J. Demmel, J. Dongarra, I. Duff, S. Hammarling, G. Henry, M. Heroux, L. Kaufman, A. Lumsdaine, A. Petitet, R. Pozo, K. Remington, and R. C. Whaley. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, 28:135–151, 2001.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[5] J. Burke-Cohen. An analysis of tower (ground) controller-pilot voice communications washington. Technical report, US Department of Transportation, Office of Research and Development, 1995.

[6] K.M Cardosi. An analysis of en route controllerpilot voice communications. Technical report, Washington, DC: US Department of Transportation, Office of Research and Development, 1993.

[7] M.A. Carlin and B.Y. Smolenki. Detection of speaker change points in conversational speech. In *Aerospace Conference, 2007 IEEE*, 2007.

[8] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[9] S. Galligan. Automatic speech detection and segmentation of air traffic control audio using the parametric trajectory model. In *Aerospace Conference, 2007 IEEE*, 2007.

[10] H. Gish, M. Siu, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991.

[11] David Guillamet and Jordi Vitria. Classifying faces with non-negative matrix factorization, 2002.

[12] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011.

[13] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[14] Hyuonsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[15] William H. Lawton and Edward A. Sylvestre. Self modeling curve resolution. *Technometrics*, Vol. 13, No. 3:617–633, 1971.

[16] A. Lechner, P Mattson, and K. Ecker. Voice recognition: software solutions in real-time atc workstations. *IEEE Aerospace and Electronic Systems Magazine*, 17:11–16, 2002.

[17] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[18] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[19] Tracy Lennertz, Judith Bürki-Cohen, and Andrea L. Sparko. Nextgen flight deck data comm: Auxiliary synthetic speech phase i. Technical report, U.S. Department of Transportation, 2012.

[20] Chin J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[21] (Nuance) Loquendo. www.nuance.com.

[22] Carol Manning, Scott Mills, Cynthia Fox, Elain Pfleiderer, and Henry Mogilka. Using air traffic control taskload measures and communication events to predict subjective workload. Technical report, Department of Transportation, 2002.

[23] Gustavo Niemeyer. Geohashing (www.geohash.org).

[24] Travis E. Oliphant. Python for scientific computing. *Computing in Science and Engineering*, 9(3):10–20, 2007.

[25] J. M. Pardo, J. Ferreiros, F. Fernandez, V. Sama, R. de Cordoba, J. Macias-Guarasa, J. M. Montero, R. San-Segundo, L. F. D'Haro, and G. Gonzalez. Automatic Understanding of ATC Speech: Study of Prospectives and Field Experiments for Several Controller Positions. *IEEE Transactions on Aerospace Electronic Systems*, 47:2709–2730, 2011.

[26] Robert Peharz, Michael Stark, and Franz Pernkopf. Sparse nonnegative matrix factorization using l0-constraints. In IEEE, editor, *Proceedings of MLSP*, volume n/a, pages 83 – 88, Aug 2010.

[27] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[28] Daniel Ramage and Evan Rosen. Stanford topic modeling toolbox (http://nlp.stanford.edu/software/tmt/tmt-0.4/), 2004.

[29] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[30] R. Clint Whaley, Antoine Petitet, and Jack J. Dongarra. Automated empirical optimization of software and the ATLAS project. *Parallel Computing*, 27(1–2):3–35, 2001. Also available as University of Tennessee LAPACK Working Note #147, UT-CS-00-448, 2000 (`www.netlib.org/lapack/lawns/lawn147.ps`).

[31] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM.