2016 IEEE Second International Conference on Big Data Computing Service and Applications

# TopicListener: Observing Key Topics from Multi-Channel Speech Audio Streams

Jing Su*, Oisín Boydell†
*Centre for Applied Data Analytics Research*
*University College Dublin*
*Ireland*
*Email: *jing.su@ucd.ie, †oisin.boydell@ucd.ie*

***Abstract*—Speech audio often encapsulates huge volumes of information which traditionally has been challenging to mine and analyse using automated methods. For example, call centres often handle many simultaneous telephone conversations between customers and call centre agents where, apart from relying on limited manual reporting by individual call centre agents, the content, themes and topics of the conversations are not analysed in any depth. In recent years there have been significant improvements in both the accuracy and cost of automated speech-to-text transcription technologies which can be applied in the call centre environment. We introduce TopicListener, which combines advanced topic modelling techniques with automatic speech transcription to identify key themes and topics across large volumes of recorded audio conversions as well as providing a novel means to explore and visualise the correlation and evolution of topics over time.**

## I. INTRODUCTION

There is vast amount of valuable information contained in speech audio (conversations, dialogue, spoken presentations, talks, commentary etc.) found across many different industries which traditionally has been nearly impossible to mine and analyse in an automated way. This is primarily due to two factors: firstly, extracting spoken words from an audio signal (automated speech-to-text transcription) has until recently been highly error prone, hard to adapt to specific subject domains, accents and languages, and expensive to deploy. Secondly, once the spoken words have been extracted as text the content often requires some form of analysis in order to further understand and extract meaningful insight. Although text analytics is a relatively mature field, such analysis of automatically transcribed speech provides additional complexities that would not usually be encountered in traditional text corpora such as high word error rate and limited structure.

An example of where such automated analysis of speech audio promises to bring huge benefits is in the call centre industry. A typical call centre fields multiple incoming calls from customers, in some cases many hundreds simultaneously, where each call is handled by a different call centre agent. Although each agent is tasked to resolve individual customer issues and queries and possibly to provide some limited manual reporting of the nature of the call there is generally no means to provide an accurate overview or summary of the overall key issues being handled by the call centre at any time. This is due to the complexity in analysing speech audio as discussed above as well as the sheer volume of calls that many of these call centres routinely handle.

The TopicListener system aims to address this challenge by leveraging recent advancements in automatic speech-to-text transcription technology which enables its deployment in high volume settings such as call centres to produce high accuracy, low cost and in near real-time transcriptions. TopicListener then focuses on applying approaches from the field of topic detection and tracking (TDT) to this transcribed content and we introduce a novel topic tracking and visualisation approach to assist the understanding and interpretation of emerging issues, topics and themes encountered in a call centre and how they evolve over time. Figure 1 illustrates a basic approach of topic detection and tracking from multichannel audio streams. Speech recognition is applied independently on each audio channel. The transcriptions are then collected as one corpus with no differentiation between channels. Ranked topics are generated from that corpus using a topic modelling algorithm. A complete topic tracking system is introduced in Figure 3. The TopicListener system features a divide-and-conquer approach over the 'big-data' challenge whereby different topic models are generated sequentially over different temporal subsets of the data. Moreover, this sequential processing mode improves user experience and assists the discovery and monitoring of topics as they evolve over time.

The paper is arranged in the following structure. In Section II-A we present the state-of-the-art in speech analytics for call centres and show that a deeper analysis of the actual content of calls across the full call centre is not being addressed. In order to meet this need, we explore recent advancements in speech-to-text transcription (Section II-B), topic modelling (Section II-C) and topic visualisation (Section II-E). However, the available technologies do not offer an off-the-shelf solution for the 'big-data' challenge in speech analysis and accomplish unsupervised key information extraction. We introduce the architecture of TopicListener and bring an end-to-end solution (Section

978-1-5090-2251-9/16 $31.00 © 2016 European Union
DOI 10.1109/BigDataService.2016.23

III) and explain the advantages of our visualisation system (Section IV). A comprehensive evaluation is then made on speech transcription (Section V-A), topic modelling (Section V-B) and visualisation (Section V-C).

## II. BACKGROUND

We first review the state-of-the-art in speech analytics solutions for the call centre industry who's limitations motivate our application of topic modelling to multi-channel transcribed speech audio. We then introduce core techniques that are integrated to deliver the TopicListener system: automatic speech-to-text transcription technology (Section II-B), topic modelling (Section II-C), model comparison metrics (Section II-D) and visualization of topics and their evolution over time (Section II-E).

### A. Speech Analytics Solutions for Call Centres

Speech analytics refers to analytics carried out over speech audio. There are a number of key approaches that are commonly used:

- Pitch, tone and talk-over analysis
- Phonetic Indexing
- Full Transcription

Pitch, tone and talk-over analysis involves analysing the audio features of speech to determine emotional aspects of the conversation such as anger, dissatisfaction, irritation, excitement etc. For example in a call centre scenario a strong indicator of caller dissatisfaction is when the caller talks over the call centre agent and this can be detected and measured by the audio signal. NICE Systems[1] produce a call centre speech analytics solution that measures customer emotion using pitch and tone analysis as well as talk-over detection. These approaches can be considered as light-weight speech analytics techniques as the actual content of the speech is not analysed in any direct way.

Phonetic indexing is where the speech audio is converted into a string of phonemes, the basic units of speech, and these are then indexed. The phonemes are not converted into actual textual output so the speech content cannot be read or mined as text; however, indexed phoneme content can be searched for specific known words or phrases. An example in a call centre speech analytics solution would be to count the number of calls where a competitor's product name was mentioned. A number of commercial speech analytics solutions use phonetic indexing as their core technology. Examples include Nexidia[2], Callfinder[3] and Univoc KWS[4]. VPI[5] provide a solution that offers conceptual search of call data - rather than restricting searches to an exact match with the search query, results for related concepts are also returned. Although exact details are not given this is likely to be achieved through query expansion techniques [1] being applied before a standard phonetic index search. The main limitation of phonetic indexing is that it can only be used for searching for words or concepts known in advance - it cannot be used for detecting new and previously unknown concepts or issues.

Full transcription attempts to produce a verbatim transcription of the speech audio into raw text. This textual content can then be mined and analysed in much more depth than the previous approaches allow. However until more recently full transcription technologies were costly to deploy and accuracy was low which meant that their use in call centre speech analytics solutions was limited. In the next section we discuss how recent advances in full transcription technology has opened the way for their more widespread use in call centre speech analytics and how our approach outlined in this paper is thus suitable for real-world application.

### B. Speech-To-Text Transcription

Full speech-to-text transcription technology has progressed enormously in recent years. One of the drivers behind this is the 'big-data' approach that has been made possible through large companies becoming involved, most notably Google and Apple. The ability to collect huge volumes of sample speech data through end user services such as voice search and manually captioned online video has enabled detailed refinement and improvement of speech transcription algorithms and the acoustic models on which they are based which has resulted in a significant increase in transcription accuracy. Previously, accurate transcription was often limited to specific niche subject areas where a system could be finely tuned to a specific vocabulary and for a specific set of speakers. With current technology the accuracy for general purpose transcription for different speakers with different languages and accents has improved, and this trend is likely to continue. The cost of large scale deployment is also coming down. This is evident with the wide scale adoption of speech recognition powered services such as Google's voice search[6], automatic caption generation for Youtube content[7] and Apple's Siri[8] personal assistant.

Of course simply extracting the textual content of audio calls is only part of what is required in order to understand and make sense of the kinds of high call volumes that are common in a busy call centre environment. The analysis of the transcribed text from multiple simultaneous calls and over time is a challenge that we address in TopicListener through the application of topic modelling techniques.

[1] http://www.nice.com/speech-analytics
[2] http://www.nexidia.com/nexidia/about_us/phonetic_search_technology
[3] http://www.mycallfinder.com/callfinder-features/phonetic-indexing/
[4] http://www.univoc.ca/english/exploration/
[5] http://www.vpi-corp.com/call-center-analysis-mining.asp
[6] http://www.google.ie/insidesearch/features/voicesearch/index-chrome.html
[7] http://googleblog.blogspot.ie/2009/11/automatic-captions-in-youtube.html
[8] http://www.apple.com/ios/siri/

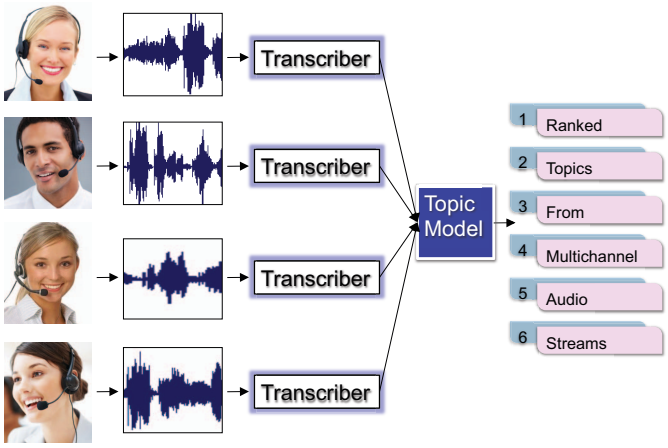



Figure 1. TopicListener application scenario: Modelling topics from multi-channel call centre audio transcriptions

### C. Topic Modelling

There has been significant research over the last ten to fifteen years in the area of topic detection and tracking (TDT), also known as topic modelling [2] [3]. This is concerned with the identification of cohesive topics or events described across multiple textual sources with these topics not being known in advance. There is also a temporal element whereby new sources might continuously become available and previously detected topics may evolve or disappear thus the tracking of topics over time is of interest. This last point is particularly relevant to the modelling of topics in transcribed call centre calls as new calls are continually being handles by call centre agents.

Popular topic modelling algorithms have evolved through Latent Semantic Analysis (LSA) [4] and probabilistic Latent Semantic Indexing (pLSI) [5] to Latent Dirichlet Allocation (LDA) [6] and Non-negative Matrix Factorization (NMF) [7]. LSA is an information retrieval approach performing dimension reduction on a document-term vector space, in order to estimate the similarities between documents or between a document and query. pLSI improved on LSA by introducing solid statistical foundation and defining a proper generative data model. LDA is a generative model which regulates the relation of documents and topics as well as topics and words with probability distributions. LDA is explained in details in Section II-C1. NMF which has been more traditionally applied for other tasks such as data exploration in bioinformatics and document clustering has also been found to be effective for topic modelling. Beyond these models, Hierarchical Dirichlet Process (HDP) [8] is a widely used hierarchical and non-parametric model for clustering multiple grouped data. HDP has advantage in estimating the number of mixture components from data. The details of HDP is introduced in Section II-C2.

The classic application of topic detection and tracking techniques is in the analysis of news reporting [9] [10]. New topics or events are continually reported across multiple news sources. These events are not known in advance so must be 'detected' or identified from the available sources, and as circumstances change and news reporting follows each event the event or 'topic' may evolve or else fade from the reporting spotlight - hence the 'tracking' aspect. TDT techniques have also been applied over streaming short text data such as Twitter [11]. There are close similarities between these application areas and call centre calls, once transcribed to text. The same issues, queries and points of discussion which can be identified through topic modelling as distinct topic often occur across multiple calls from different customers. Many of these are likely to occur around the same time period in response to external events such as when a new service or product is rolled out which results in many customers experiencing similar experiences around the same time. Thus topic detection and tracking is very applicable for a novel call centre analytics approach.

*1) Latent Dirichlet Allocation (LDA):* Blei et al. [6] introduced Latent Dirichlet Allocation (LDA) as a generative probabilistic model for collections of discrete data such as text corpora. LDA regulates the probabilistic distributions between document, topic and word and it is an unsupervised learning model. Previous topic modelling approaches, such as the mixture of unigrams model can only extract a single topic per document [12] whereas LDA defines a Dirichlet distribution between a document and multiple topics. This is a more flexible model in practice. Specifically, LDA assumes that the terms in corpus are generated by the following process:

(1) For the $k$–th topic
  Sample words $\beta_k \sim Dir_V(\eta)$
(2) For the $d$–th document $w_d$:
  sample topic proportion $\theta_d \sim Dir(\alpha)$
  For word $w_n, n \in \{1,2,...,N\}$
    Sample a topic $z_{d,n} \sim Mult(\theta_d)$

Sample a word $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

This process can be represented by a joint distribution in the form:

$$P(z,w,\theta) = P(\theta_d|\alpha)\prod_{n=1}^{N}[P(z_{d,n}|\theta_d)P(w_{d,n}|z_{d,n},\beta)] \quad (1)$$

The objective is to find the posterior distribution of latent variables:

$$P(z,\theta|w,\alpha,\beta) = \frac{P(z_{d,n},w_{d,n},\theta_d|\alpha,\beta)}{P(w_{d,n}|\alpha,\beta)} \quad (2)$$

Since this distribution is not tractable, variational inference is introduced for approximation [6].

*2) Hierarchical Dirichlet Process (HDP):* Hierarchical Dirichlet Process (HDP) [8] is a widely used hierarchical and non-parametric model for clustering multiple grouped data. HDP assigns a Dirichlet process for each group of data and the Dirichlet processes for all groups share a base distribution. In the trials of topic modelling with LDA, a major problem is how many topics that a corpus of texts has. HDP approaches this problem by estimating the number of mixture components from data.

HDP features a two level Dirichlet process in which a base layer Dirichlet distribution is used to sample corpus-wise topic distribution and a second layer Dirichlet distribution is used to sample document-wise topic distribution. The second layer topic distribution $G_j$ shares a common distribution of $G_0$.

$$G_0|\gamma,H \sim DP(\gamma,H) \quad (3)$$
$$G_j|\alpha_0,G_0 \sim DP(\alpha_0,G_0) \quad (4)$$

where an explicit definition of DP is given by Sethuraman [13] as Equation 7, an *stick-breaking construction* approach. The stick-breaking construction is based on independent sequences of i.i.d. random variables $(\pi'_k)_{k=1}^{\infty}$ and $(\phi_k)_{k=1}^{\infty}$ [8].

$$\pi'_k|\alpha_0,G_0 \sim Beta(1,\alpha_0) \quad (5)$$
$$\phi_k|\alpha_0,G_0 \sim G_0 \quad (6)$$

then a random measure $G$ can be defined as

$$\pi_k = \pi'_k\Pi_{l=1}^{k-1}(1-\pi'_l) \quad (7)$$
$$G = \Sigma_{k=1}^{\infty}\pi_k\delta_{\phi_k} \quad (8)$$

$\delta_\phi$ is a probability measure concentrated at $\phi$. Sethuraman [13] showed that G as defined in this way is a random probability measure distributed according to $DP(\alpha_0,G_0)$.

### D. Model Comparison Metrics

In order to assess the discrepancy of two topics in adjacent time windows and locate similar topics which allows TopicListener to present the occurrence and evolution of topics over time, we need a metric to quantify topic similarity. In this study we follow Greene's approach for measuring topic model agreement [14].

*1) Term Ranking Similarity:* The output of either an LDA or HDP generated topic model is in the form of a ranked list of $k$ topics, each topic consisting of a ranked list of terms. We can denote a topic list as $S = \{R_1,...,R_k\}$, where $R_i$ is a topic with rank $i$. An individual topic can be described as $R = \{T_1,...,T_m\}$, where $T_l$ is a term with rank $l$ belong to the topic. In order to assess the similarity of two topic sets $S_x$ and $S_y$, we need to evaluate the similarity of the individual topics between $S_x$ and $S_y$ first.

Kendall introduced a rank correlation measure, Kendall's $\tau$ [15], as a metric for comparing two ranked lists. $\tau$ is a scalar value ranging from -1 to 1, where 1 means two ranked lists are identical and -1 means one list is in the reverse order of the other. $\tau = 0$ means only 50% items in two lists match in order. But in order to be comparable with the results of previous topic stability analysis work, here we use Jaccard index [16] as a fundamental metric to compare the common terms between two topics.

Jaccard index only compares the number of identical items in two sets, neglecting any ranking order. But in a topic model, at least one produce either using LDA or HDP, the top terms weigh more in defining the characters of topic than lower ranked terms. We apply Average Jaccard (AJ) similarity [14], as a top-weighted version of the Jaccard index to accommodate ranking information. AJ (Equation 9) calculates the average of the Jaccard scores between every pair of subsets of $d$ top-ranked terms in two lists, for depth $d \in [1,t]$.

$$AJ(R_i,R_j) = \frac{1}{t}\sum_{d=1}^{t}\gamma_d(R_i,R_j) \quad (9)$$

where

$$\gamma_d(R_i,R_j) = \frac{R_{i,d} \cap R_{j,d}}{R_{i,d} \cup R_{j,d}} \quad (10)$$

In Equation 9, $R_{i,d}$ is the head of list $R_i$ up to depth $d$ and $\gamma_d$ is a symmetric metric within the range [0,1]. Therefore each term in a ranked list is weighted by its rank in a decreasing order and the top terms are more influential to an AJ score.

### E. Visualisation

Visualisation of topic models is one essential aspect of delivering topic modelling outputs [17] [18] [19] [20].

Chaney and Blei [17] present a system to organise, summarise, visualise, and interact with a corpus, in which the system is built with a fitted topic model. The system

interface works in two ways, featuring topic pages and document pages. A topic page has three columns. The left column lists the terms of a topic with the order of topic-term probability. The centre column lists documents covered by the current topic and the documents are ordered by inferred topic proportion. The right column features a list of related topics. On clicking a document name from the topic page, a corresponding document page is shown in detail, alongside a list of related topics and a list of related documents.

Chaney and Blei's approach offers a systematic way in exploring topics in a high volume corpus or articles, such as Wikipedia. Documents related to a topic are collected and ranked on one page. Topics related to a document are also sorted and easily accessible. Readers can trace relevant information in a convenient way. However, this visualisation approach indicates the importance of a document or a topic only in ranking, instead of a direct-viewing graphic representation. Moreover, the whole corpus is processed in a topic model and the topic structure is static. When we opt to display the evolution of topics in time sequence, such visualisation approach is not satisfying.

Malik et al. [19] introduce TopicFlow, a time sequenced topic visualisation tool on Twitter. TopicFlow builds LDA topic models over batches of tweets which are collected within a selected time interval. In LDA models, the default number of topics is 15. Both the duration of time intervals and the number of topics are adjustable in order to achieve proper granularity of topic modelling. Afterwards, TopicFlow employes a topic alignment step to visualise the correlations between adjacent topic models. Cosine similarity is used to compute the similarity of each pair of topics in adjacent topic models.

We present a new visualisation for topics and how they evolve over time in our TopicListener system. In the next section we describe our system in detail including our novel visualisation approach.

## III. SYSTEM DESCRIPTION

In Section II we reviewed a series of technologies related and contributive to speech recognition, topic mining and visualisation from speech audio sources. However, these approaches each solves a single challenge and there is no trials for an end-to-end solution yet. In this section, we carry out topic model selection (Section III-A) and introduce the design of the TopicListener system (Section III-B). TopicListener embodies a systematic approach for topic model generation over audio streams, especially on multi-channel call centre recordings.

### A. Topic Model Selection

The LDA model is one of the most popular topic modelling approaches (Section II-C1). However, we need to select a proper number of topics for LDA before modelling a target corpus. In call centre applications, TopicListener is expected to process unknown corpora consecutively and automatically. Therefore we are looking for an unsupervised approach to determine an appropriate number of topics for a topic model on a new corpus. Greene et al. [14] applied a stability analysis approach to determine the proper complexity level for NMF topic models. In this experiment we apply the proposed stability metric, Hungarian agreement score $H$, to evaluate the similarity of LDA models generated from a whole corpus and a portion of that corpus.

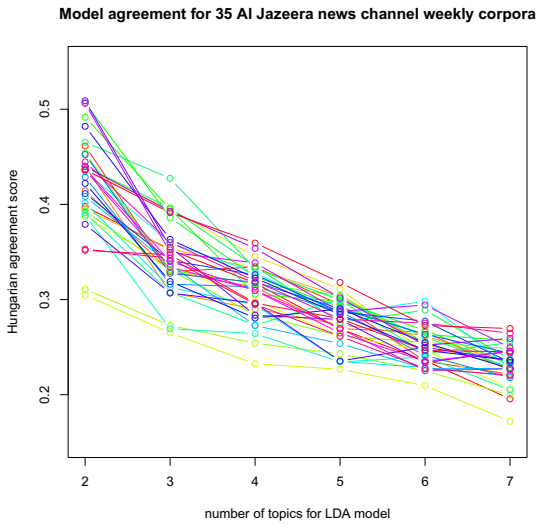



Figure 2. Topic model agreement scores of LDA models in 35 weekly Al Jazeera transcription corpora, with $k$ ranging from 2 to 7

We used a dataset consisting of transcripts of audio commentary from news report videos from the Youtube Al Jazeera English channel. This is a comprehensive international news channel with the content including politics, military, economy, sports, education, etc., and for our experiments we collected 11,120 video documentaries that were automatically transcribed by Google's automated speech transcription engine.

We then split this dataset into 35 weekly sub-corpora. On each sub-corpus $C_i$ (containing on average over 60 documents), we randomly select 80% documents as a candidate sub-corpus $C'_i$. LDA models are trained with $k$ ranging from 2 to 7 on both $C_i$ and $C'_i$. Then $H_{i,k}$ measures the similarity of LDA models $T_{i,k}$ and $T'_{i,k}$. In Figure 2 we observe Hungarian agreement scores $H_{i,k}$ against $k$ in 35 weekly sub-corpora, among which $k = 2$ matches the highest $H_{i,k}$ in almost every week. This means a very simple topic model is always selected for a highly diverse news corpus. Thus automatic model complexity selection on LDA is not suitable for the news corpora. Instead, for TopicListener, we apply the HDP model (Section II-C2) using Heinrich's [21] implementation. The training iteration is 100 for each weekly sub-corpus.

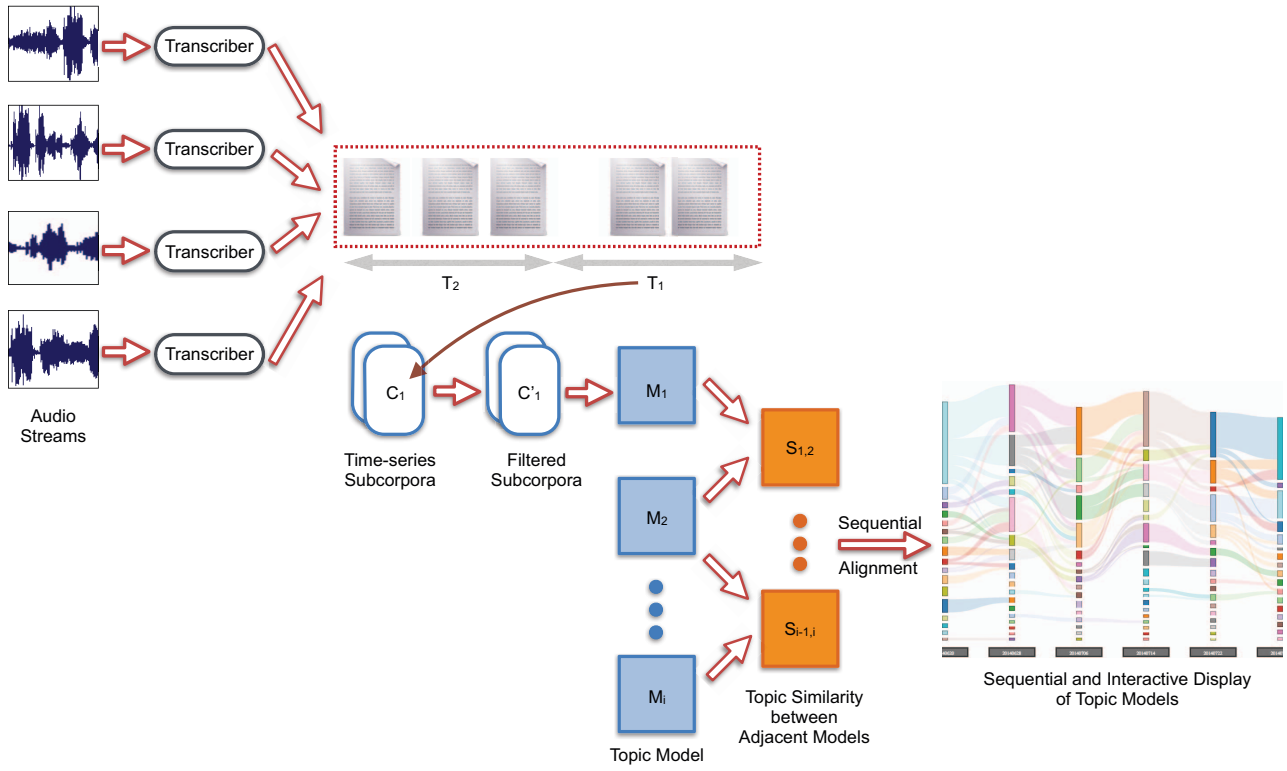



Figure 3. TopicListener system architecture: a structural approach in corpus alignment, topic modelling, topic similarity measurement and topic model visualisation

### B. *TopicListener System Architecture*

Figure 3 illustrates the architecture of TopicListener. In this image, the system works from left to right featuring five components of speech transcription, time-series sub-corpora generation and filtering, topic modelling, topic similarity measurement and visualisation. The complete system of TopicListener is developed from a basic idea of topic modelling over multi-channel speech recordings (Figure 1). Among many variations in between, the major difference is time based sampling, modelling and visualisation mode.

In Figure 3, there are four input audio streams. In the call centre applications, these represent different call centre agents holding customer calls. Individual audio segments (calls) are collected within a fixed length time interval $T$ across all streams. Collections of audio clips collected during $T_1$ are transcribed separately into texts and they compose a sub-corpus $C_1$. Therefore, information from different audio streams is collected indiscriminately and is sorted into a single sub-corpus. The system is scalable to more audio channels. An important requirement is that a document of speech transcription must be accompanied with its timing labels.

Here we highlight a filtering step prior to topic modelling in which stop words are removed from $C_1$. A stop words list needs to cover generic English stop words as well as domain specific stop words. A properly designed stop words list notably improves topic model coherence.

Following the corpus filtering step, a HDP model $M_1$ is trained over the filtered sub-corpus $C_1'$. In the same way, $M_2$ is trained over voice of 4 channels in $T_2$. Users can tell stories easily from the keywords of a single topic, but it is challenging to spot all similar topics between $M_1$ and $M_2$. In order to trace and visualise topic evolution, we generate topic similarity matrix $S_{1,2}$ between two adjacent topic models $M_1$ and $M_2$. The metric of pairwise topic similarity is Average Jaccard similarity (Section II-D1).

The TopicListener system works in an incremental scheme. When the latest speech source $T_i$ is available, it is taken as an independent input and is processed in a pipeline. The incremental mode of data processing and visualisation is especially convenient for call centre applications. Users can read the latest topics alongside historical topics while topic models iteratively process new data on daily or hourly basis.

The incremental scheme is not only beneficial for users, but also a divide-and-conquer approach for computation. The total corpus of audio recordings from a call centre can be huge in volume, which is a 'big-data' challenge. Moreover, speech recognition and topic modelling both are computation intensive tasks. It is demanding to process a Gigabyte level

repository of multi-channel recordings. On the contrary, an incremental processing scheme is designed to process a smaller size dataset while waiting for the generation of next batch of data. The workload in a fixed time interval is defined by the number of speech channels and the length of the time window. This is a much easier task.

On the right side of Figure 3 there is a snapshot of TopicListener user interface. This UI is designed to display topic models in a sequential order, and show the trend of topic evolution. In principle, the graphical representation of topic evolution trends is a key factor for human perception and understanding on topic outputs. More details are explained in the next section.

## IV. TOPIC MODEL VISUALISATION

### A. Visualisation Design

Topic visualisation is a key component of TopicListener in order to present time sequenced topic modelling output in an intuitive and objective approach.

The topic visualisation approach of TopicListener is inspired by Sankey diagrams [22]. Although there is similarity with the approach of Malik et al. [19], TopicListener visualisation has significant differences in the metrics which control the size of nodes and measure the similarity of topics.

The objectives and features of TopicListener visualisation include:

- In each topic model, highlight the major topics.
  - The size of a topic is determined by the probability of a topic in topic model, instead of the number of related documents.
  - Major topics are larger in node size.
  - Topics of one topic model are distinguished with different colours.
- Clearly show the correlations between two topics of two adjacent topic models.
  - Topic correlations are distinguished by colours.
  - The links from the same topic are identified with same colour.
- Clearly show the emergence of new topics in time sequence.
  - No incoming correlation links on a new topic.
- Clearly show the ending of a series of correlated topics.
  - No outgoing correlation links from an ending topic.
- Clearly show the trend or evolution of correlated topics in time sequence.
- Clearly show the standalone topics which has little correlation with previous and following topics.
  - No incoming or outgoing correlation links on a standalone topic.
- Easily explore the keywords in each topic.
  - Topic keywords are displayed as word cloud in an extra text area when mouse is on a topic node.
- Easily navigate and locate topics from a time period.
  - Topic nodes are draggable vertically so as to allow a clear view of the links.
  - There is a horizontal slider for navigation along timeline.

We explain an example of topic model visualisation in Figure 4.

### B. Visualisation Demonstration

Based on our one-to-one topic similarity measure, we have a matrix of topic agreement scores. This matrix covers tens of observation windows (sub-corpora). We present a user interface for visualising topic flows in sequence.

Figure 4 shows an example of topic flow visualisation. Each column of nodes (blocks) represents topics from one topic model, which are extracted from a weekly sub-corpus. In each column, each node stands for one topic. When a node is clicked on, the top terms describing a topic are displayed on the right side pane as a word cloud. In a word cloud each word has a different font size. The most popular word in a topic takes the biggest font and the remaining keywords are sorted with decreasing fonts.

The colours of nodes are different only for the purpose of separation. However, the height of each node is scaled to the weight of a topic in the topic model. The curved lines connecting pairs of topics in consecutive windows indicate a topic similarity higher than a threshold. With this interface, it is convenient to trace along curves for similar stories occurring in sequential order. The emerging topics and diminishing topics are also easily located.

In Figure 4 we highlight three nodes to illustrate how it works in the UI. Node 1 is an emerging topic that occurred in the week of 2014-03-20, and it tells the story about "refugees, security, kenya" etc. Node 1 is related to only one topic in the following week, which is node 2. The topic of node 2 covers "refugees, lebanon, syrian" etc, which are related to refugee problems but they happen in different countries. The topic of node 3 covers "syrian, city, forces" etc which is no longer about refugee but military actions related to Syria. Node 3 has ongoing stories in the corpus of the following week.

Examining the links between these three nodes, we can see that related topics or stories are correctly labelled and linked together. In this case, the correctness of proper links attributes to average Jaccard similarity (Section II-D1). Another merit of topic linking is to indicate emerging topics as well as ending topics. For example, the topic of node 1 is not popular in the previous week, so it can be taken as an important breaking news.

## V. EVALUATION

The TopicListener system incorporates a series of techniques to process speech audio, generate topic models and visualise topic models in time sequence. The effectiveness

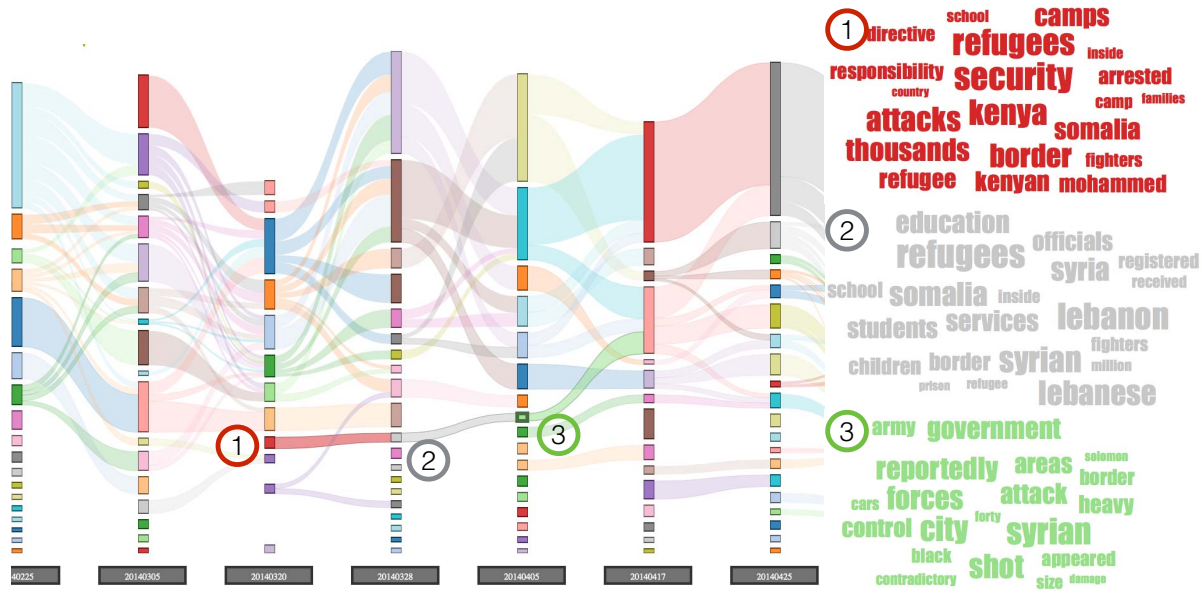



Figure 4. TopicListener visualisation with topics captured in March and April 2014 from automatic transcripts of the Al Jazeera news channel

and reliability of final outputs are determined by numerous factors including speech recognition accuracy, topic modelling efficiency and the usability of visualisation. In this section we present ways of evaluation on TopicListener and emphasise on system reliability examination.

### A. Speech Transcription Evaluation

Text corpora generated from automatic speech recognition inevitably contain transcription errors. The level of transcription errors varies according to the audio quality of speech recordings as well as the capability of speech recognition engines. In section II-B we address the advances of automatic speech recognition brought by the 'big-data' approach. Consequently, we opt to use Google's automatically generated news channel captions as the corpus for our experiments.

Google's automatic captions offer a convenient approach to collect large quantity of data. However, Google does not offer an accuracy score of the automatically generated captions, and it is overwhelming to manually verify the accuracy of our corpus which covers over 300 hour news recordings. After inspecting a number of captions against their corresponding audio clips, we find the transcriptions are reliable and most of the named entities are spelled correctly. This feature can be attributed to the voice of professional news reporters and high quality studio recording operations.

However, the TopicListener system is expected to retrieve topics from multi-channel call centre recordings. In that scenario noise and cross-talk seriously challenge speech recognisers. Would topic modelling output from noisy text corpora be reliable? We evaluate the stability of topic model against transcription errors in the next section.

### B. Topic Model Stability Evaluation

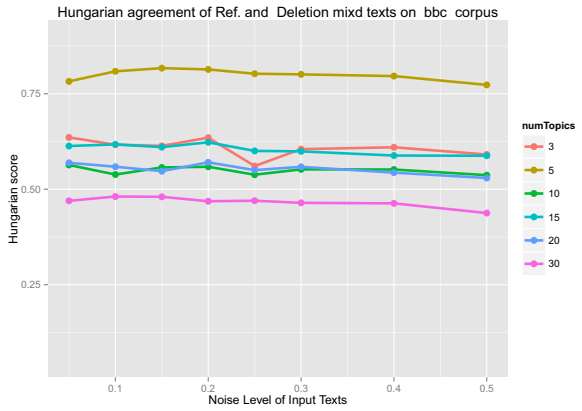
As discussed in Section III-A, we select HDP model over LDA to avail the convenience of unsupervised topic modelling. Since the model complexity is controlled by HDP, a major challenge of topic modelling is the stability of models against textual noise in the input corpus. In call centre applications, text corpora are generated from automatic speech recognition where transcription errors are inevitable. In some cases the errors can be serious. Therefore the objective of this evaluation is to test the stability of topic models over noisy corpora.

We design an evaluation method for topic model stability (or robustness). The idea is to run topic models over reference corpus and noisy corpus and compare the similarity of output topic models. In order to make the textual noise controllable, we introduce artificial word errors including deletion, insertion and replacement. These errors are analogous to word error rate (WER) [23]. *Deletion* errors are introduced by randomly remove 0% to 50% terms from an article and the term selection is based on uniform distribution. *Insertion* and *Replacement* errors are introduced by adding 0% to 50% random terms from a list of frequent English words with 7726 entries[9], and the sampling probability is based on term frequency.

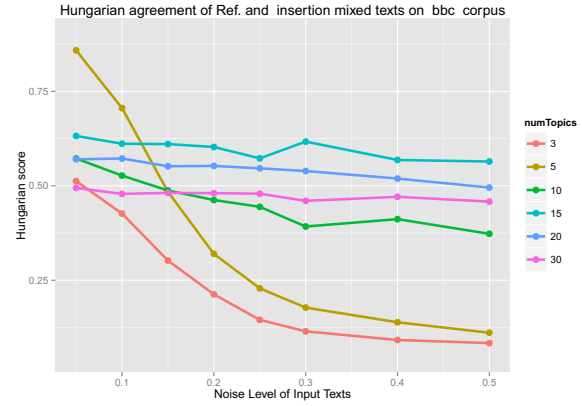
LDA topic model is used to evaluate model robustness w.r.t. model complexity (number of topics)[10], and the similarity measure of topic models is Hungarian agreement score

[9]http://ucrel.lancs.ac.uk/bncfreq/flists.html

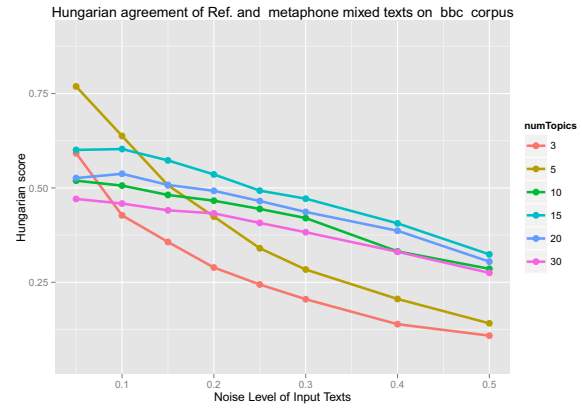
[10]Since HDP model selects the model complexity automatically, we opt to use LDA with the ease of manual setting




(a) Deletion errors




(b) Insertion errors




(c) Replacement errors

Figure 5. LDA model Hungarian agreement scores with various levels of *Deletion* errors (a), *Insertion* errors (b) and *Replacement* errors (c) in the *bbc* corpus.

[14]. In *bbc* corpus[11] [14], the reference topic number is 5. In Figure 5 we see that topic models with the reference number of topics achieve the highest stability scores when noise level is relatively low (below 15%). Surprisingly, topic models are stable with *Deletion* errors up to 50%. From this evaluation we assume that HDP, as an extension model of LDA, shares similar performance against different types of textual noise, and is especially robust with deletion errors.

Topic modelling over noisy transcription sources is trustworthy if we can control the noise level to be low or avoid insertion or replacement types of errors. As a general suggestion to transcriber configuration (e.g., acoustic model and language model) for the task of topic modelling, we recommend to remove uncertain outputs of transcription because deletion errors do not influence much on topic modelling than an spurious term (insertion or replacement error).

### C. Visualisation Evaluation

In Section IV-B we demonstrate TopicListener visualisation with topics captured from automatic transcripts of a news channel. Although it is straightforward to read topic keywords from the interface, it is difficult to have an objective evaluation of the visualisation system. We therefore propose a subjective approach of evaluation. The TopicListener system was tested over automatic transcripts of call centre recordings from the financial servicing industry and the visualisation output was evaluated by call centre professionals. The response was that the visualization was a useful tool that assisted the understanding and exploration of the extracted topics.

## VI. CONCLUSION

Among vast quantities of speech audio resources, especially multi-channel call centre recordings, people are overwhelmed by the quantity of information and the limited approaches currently on offer to analyse the content. Although automatic speech transcription technologies bridge the gap between linear access to audio sources and non-linear access to text information, there is still a significant challenge in automatic and effective text information retrieval. Topic modelling is a widely applied approach in text summarisation and topic extraction. In this study we focus on the challenge of topic modelling on unsupervised audio stream monitoring, and propose a robust topic modelling tool in solving this problem. Beyond addressing the challenge of audio volume, we also highlight the importance of timing in data processing and topic modelling. An extracted topic delivered along with its time of occurrence is essential for users. Therefore, sub-corpus organisation, topic modelling and visualisation are all based on a predefined time interval.

[11]The *bbc* corpus consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005, specifically business, entertainment, politics, sport and technology.

Another advantage of this work is the sequential and interactive user interface which illustrates the evolution of topics in an intuitive way. The TopicListener system can also be used to explore topics in audio/video news documentaries.

We prefer to summarise our approach, TopicListener, as a general purpose topic monitoring tool over time sequenced audio sources. It is not a traditional categorisation or classification tool where the topics are defined in advance. TopicListener detects the core topics automatically and it attempts to determine the appropriate number of topics in data. Moreover, an innovative user interface is presented in this work. Users can easily track the evolution of topics, understanding the change of topic content and popularity. We look forward on more applications of TopicListener with the rapidly increasing volumes of speech audio sources that are becoming available.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support of Enterprise Ireland and IDA Ireland for funding this research.

## REFERENCES

[1] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1:1–1:50, Jan. 2012.

[2] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 37–45.

[3] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 28–36.

[4] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, no. 25, pp. 259–284, 1998.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[7] S. Arora, R. Ge, and A. Moitra, "Learning topic models - going beyond SVD," *CoRR*, vol. abs/1204.1956, 2012.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[9] M. Mori, T. Miura, and I. Shioya, "Topic detection and tracking for news web pages," in *IEEE/WIC/ACM International Conference on Web Intelligence, 2006.*, pp. 338–342.

[10] P. Kim and S. H. Myaeng, "Usefulness of temporal information automatically extracted from news articles for topic tracking," vol. 3, no. 4, pp. 227–242, Dec. 2004.

[11] D. Ramage, S. T. Dumais, and D. J. Liebling, "Characterizing microblogs with topic models," in *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[12] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Mach. Learn.*, vol. 39, no. 2-3, pp. 103–134, May 2000.

[13] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[14] D. Greene, D. O'Callaghan, and P. Cunningham, "How many topics? stability analysis for topic models," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, 2014, pp. 498–513.

[15] M. Kendall, "Rank Correlation Methods," *Charles Griffin & Company Limited*, 1948.

[16] P. Jaccard, "The distribution of flora in the alpine zone." *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[17] A. J. Chaney and D. M. Blei, "Visualizing topic models," in *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*, 2012.

[18] N. Günnemann, M. Derntl, R. Klamma, and M. Jarke, "An interactive system for visual analytics of dynamic topic models," *Datenbank-Spektrum*, vol. 13, no. 3, pp. 213–223, 2013.

[19] S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, and B. Shneiderman, "Topicflow: Visualizing topic alignment of twitter data over time," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 720–726.

[20] D. Ganguly, M. Ganguly, J. Leveling, and G. J. Jones, "Topicvis: A gui for topic-based feedback and navigation," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, USA: ACM, 2013, pp. 1103–1104.

[21] G. Heinrich, "Infinite LDA - Implementing the HDP with minimum code complexity," arbylon.net, Tech. Rep., 2011. [Online]. Available: http://arbylon.net/publications/ilda.pdf

[22] W. L. O'Brien, "Preliminary investigation of the use of sankey diagrams to enhance building performance simulation-supported design," in *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, ser. SimAUD '12. San Diego, CA, USA: Society for Computer Simulation International, 2012, pp. 15:1–15:8.

[23] G. Saon, B. Ramabhadran, and G. Zweig, "On the effect of word error rate on automated quality monitoring," in *Spoken Language Technology Workshop, 2006. IEEE*, Dec 2006, pp. 106–109.