© Copyright HDSKD
All Rights Reserved

ArPod2.0: A Novel Arabic Podcast Dataset for Spoken Topic Identification

Khaled Lounnas, Mohamed Lichouri, Mourad Abbas

Computational Linguistics Departement- CRSTDLA. Algiers-ALGERIA

LCPTS-USTHB University. Algiers-ALGERIA
HCLA. Algiers-ALGERIA

DOI: 10.5281/zenodo.14810807

***Abstract*—This study introduces a new corpus designed to explore spoken topics in Arabic. Despite the growing interest in building new resources for Arabic, to our knowledge, there is no existing corpus designed specifically for recognizing the topics discussed. We assembled a substantial collection of Arabic podcasts, known as ArPod2.0, covering diverse subjects such as Business, Health, and Technology. This corpus comprises over 30 hours of speech, featuring both standard Arabic and various dialects from regions including Saudi Arabia and Egypt. Careful attention was given to curating this dataset to facilitate computer-based analysis of speech. Our focus was on six main topics, and we employed advanced computer techniques to analyze the data. Our findings highlight the efficacy of one such technique, the Multilayer Perceptron (MLP), which demonstrated exceptional performance in accurately discerning topics, even amidst mixed content.**

Keywords— Modern Standard Arabic (MSA), Dialectal Arabic, Speech Corpus, Topic, Multilayer Perceptron Models (MLP), Support Vector Machines (SVM).

1 Introduction

Spoken Topic Identification (STI) plays a vital role in Natural Language Processing (NLP), enabling machines to extract semantic information from spoken audio recordings and uncover the central themes within them. STI finds applications in various domains, including content recommendation systems, voice-activated assistants, and sentiment analysis. Understanding the context and challenges of STI sheds light on its significance and potential advancements.
In practical terms, STI enhances machines' ability to comprehend and respond to user queries effectively. For instance, in voice-activated assistants, accurately discerning the user's spoken topic is crucial for delivering relevant responses. Similarly, content recommendation systems leverage STI to identify topics within spoken content, improving the precision of suggested materials and enhancing the user experience. Despite its promising applications, STI faces obstacles:

- **Deciphering Complexities**: Accurately deciphering subject matter from speech can be intricate due to variations in accents, speech patterns, and background noise. The scarcity

© Copyright HDSKD
All Rights Reserved

of comprehensive datasets and linguistic resources specific to Arabic topics exacerbates this challenge.

• **Navigating Nuances**: Ambiguities and subtle contextual nuances inherent in spoken language complicate the task. Handling multilingual scenarios and recognizing topics in dynamic conversations pose additional hurdles.

While the research community has produced numerous spoken corpora in various languages, Arabic datasets remain notably scarce. This paper addresses this gap by introducing ArPod2.0, a substantial Arabic speech corpus exceeding 30 hours. This dataset is meticulously tailored for STI tasks and emphasizes diverse Arabic dialects, including Saudi, Modern Standard Arabic, Syrian, Lebanese, and Egyptian, while also incorporating English. ArPod2.0 is a significant contribution to research on spoken topics and represents the inaugural creation of a multilingual Arabic spoken dataset sourced from web podcasts.

The paper is structured as follows: Section II outlines the spoken topic corpora background. In Section III we present the overall framework, emphasizing two key components the collected dataset and its acquisition process, as well as its analysis. Subsection IV discusses the Spoken Topic Identification module. Sections V and VI explore the experimental setup, the results obtained, and the discussion of findings. The paper concludes in Section VII.

2 Related work

STI relies on the creation of large, natural speech corpora that are then analyzed computationally to extract thematic content. However, a review of the field reveals a significant gap in research focusing on Arabic STI. The majority of existing works explore STI in languages like English and French (Pappagari et al.).

A. Fisher based English spoken topic Corpus

The research discussed in (Park and Glass 2008), (Liu 2017), (Sung 2020), (Huijbregts 2011), (Kesiraju 2017) explores the use of unsupervised acoustic unit discovery for identifying topics in spoken audio documents. The majority of these studies focus on conducting experiments with the Fisher phase 1 English corpus. Specifically, this corpus consists of conversational telephone speech recordings, where each recording represents a different topic discussed during a telephone conversation. The authors provide comprehensive statistics regarding the distribution of recordings across topics in both the training and test sets. These statistics offer valuable insights at the intersection of machine learning, spoken language technologies, and topic identification. The findings have substantial potential for practical applications in the fields of machine learning and spoken language technologies. A novel method called Maximum Conditional Entropy (MCE) training technique was investigated by (Hazen 2008), (Hazen 2011a), (Hazen 2011b) to enhance topic identification in spoken audio documents. Through experiments conducted on the English Phase 1 subset of the Fisher Corpus, which amounted to 553 hours, considerable improvements in accuracy were observed in identifying topics. The research highlights the effectiveness of MCE in tackling the unique challenges posed by spoken audio, underscoring the importance of specialized datasets like the Fisher Corpus for refining topic identification systems. Regarding the same corpora, (Wintrode 2011) investigates the decline in topic identification

© Copyright HDSKD
All Rights Reserved

accuracy in conversational speech when diverse semantic topics are merged into a broad category. The study introduces techniques leveraging document clustering and Latent Dirichlet Allocation (LDA) to uncover underlying semantic topics, resulting in a performance improvement of up to 20% compared to classifiers trained on a single-category label. May et al. (May 2015) analyzed the performance of several feature learners in identifying subjects in speech and text data. The study examines the performance of several representations, including LDA, SAGE, mi-vector, LSA, and tfidf, over many dimensions. To conduct this research, the authors use Part 1of the Fisher English speech corpus, which includes audio recordings and transcriptions of phone calls. Their study focuses on topic identification subgroups, with 1374 training talks and 686 test conversations, each comprising two speaker specific documents.

B. Autonomous Operator of Parisian Transports (RATP)

The Spoken Dialogues Corpus, extracted from the customer care service (CCS) of the RATP Paris transportation system as part of the DECODA project, contains human-to-human telephone conversations automatically transcribed. With 1,242 conversations and a total duration of approximately 74 hours, this corpus forms the core dataset for evaluating the QMLP (Quantum Machine Learning Platform) in identifying conversation themes. The research by Bost et al. (Bost 2015) is noteworthy since it uses this corpus to present an automated analysis approach for telephone conversations between consumers and customer care representatives in the Paris public transportation system. The major goal is to acquire statistical insights into customer concerns in order to improve user happiness and prioritize essential remedies. The study utilizes a spoken topic dataset consisting of 1,654 telephone calls retrieved from the contact center of Paris' public transportation service (RATP). This dataset, which has been manually transcribed and labeled thematically, serves as the foundation for the research investigations. The limitations of conventional Multilayer Perceptron (MLP) models in Spoken Language Understanding tasks are examined by Parcollet et al. (Parcollet 2016) in their work, emphasizing their dependence on oversimplified word- or topic-based features. The authors present an original method called QMLP, which uses Quaternion algebra to capture complex dependencies inside document characteristics, to address this problem. By means of testing on real human spoken conversations, the results demonstrate the advantage of QMLP in conjunction with newly developed document properties. This strategy remarkably improves accuracy by 3% over the original Quaternion-based features and by 2% over traditional MLPs. Furthermore, the QMLP design shows efficiency by achieving impressive accuracy levels with fewer iterations. In the domain of audio-based topic analysis, there's a noticeable gap in research attention when it comes to spoken themes in Arabic (Abusulaiman 2017). Despite the growing interest in this field worldwide, the majority of research tends to focus on English and French-based topics, leaving a substantial void in our understanding of spoken Arabic audio. To overcome this gap, we investigated a recently generated spoken corpus customized exclusively for Arabic themes. Our key contribution is the creation of an Arabic spoken corpus especially intended for subject identification.

© Copyright HDSKD
All Rights Reserved

3 Data Collection and Preparation

In this section, we explain how we collected and prepared the data for ArPod2.0, a dataset designed to help identify topics in spoken Arabic content. Given the need for high-quality, standardized data in Arabic language research, we carefully selected and processed this dataset to support advanced and reliable system.

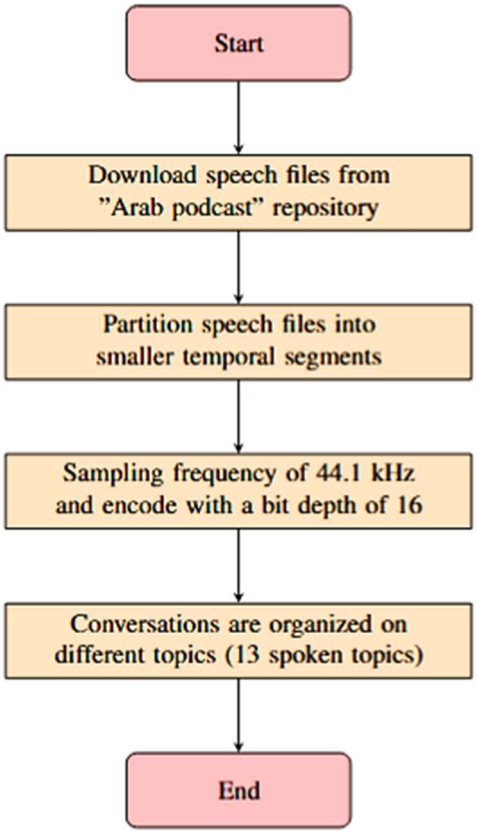

Fig. 1. Diagram for ArPod2.0 data collection and organisation.

As shown in Figure 1, ArPod2.0 is made from a range of Arabic podcasts. We used MKVToolNix to split these recordings into detailed segments. This method allowed us to capture speech from different speakers, dialects, and topics. ArPod2.0 contains over 30 hours of content and offers several important features:

1) Diverse Language Coverage: ArPod2.0 includes various Arabic dialects, such as Saudi, Syrian, Lebanese, and Egyptian, along with Modern Standard Arabic and some English. This variety addresses the lack of multilingual resources for spoken topic identification research.

© Copyright HDSKD
All Rights Reserved

2) Detailed Speech Analysis: By segmenting the recordings precisely, ArPod2.0 helps study subtle speech details and context changes. This improves the accuracy of models designed to identify topics in spoken language.

3) Real-World Conversations: Since the dataset is based on podcasts, it captures natural and engaging conversations. This provides realistic and high-quality data, ideal for training systems that can handle real-world language use.

A. Data Collection

We collected more than 30 hours of speech from the ”Arab podcast” website1. This collection includes different types of Arabic, like Modern Standard Arabic (MSA) and dialects from places like Saudi Arabia, Syria, Egypt, and Lebanon. Sometimes, you’ll also hear English in the recordings. Some parts of the recordings from Egypt, Lebanon, and Saudi Arabia mix in English words. This might affect how well our system works compared to the rest of the recordings.
To make it easier for our system to learn and work better, we split the long recordings into smaller parts, each about five minutes long. We used a tool called MKVToolNix GUI version 31.0.0 to do this. All the recordings have the same quality, with clear sound and a natural tone. In each recording, you’ll hear conversations between two or more people, including both men and women. We’ve summarized all the topics and how long they are in Table I.

B. Analysis of ArPod2.0 Characteristics

ArPod2.0 encompasses a wide range of topics used for spoken topic identification in Arabic content. Table I provides an overview of these topics and their durations. On average, each topic spans about 2 hours and 18 minutes. The ”General” topic is the longest at 3 hours and 42 minutes, while ”Health” is the shortest at 1 hour and 39 minutes.

TABLE I DESCRIPTION OF ARPOD2.0 DATASET: DURATION AND PERCENTAGE DISTRIBUTION PER TOPIC

Topic	Duration (Hours)	Percentage
Business (BU)	02:15:14	7.04%
Cinema (CI)	01:45:44	5.79%
Community (COM)	01:50:15	6.06%
Gaming (GAM)	02:53:20	9.47%
General (GE)	03:42:08	12.11%
Health (HE)	01:39:29	5.44%
Literature and Education (LAE)	02:15:20	7.06%
News and Media (NAM)	02:40:18	8.81%
Personal (PER)	02:08:27	7.04%
Science (SCI)	02:15:18	7.05%
Self-improvement (SE)	02:25:10	7.93%
Sport (SP)	02:14:54	7.01%
Technology (TECH)	02:19:18	7.61%
Total	30:23:18	100.00%

© Copyright HDSKD
All Rights Reserved

The distribution of the dataset across different topics is as follows:

- Business: 7.04%
- Cinema: 5.79%
- Community: 6.06%
- Gaming: 9.47%
- General: 12.11% and Health: 5.44%
- Literature and Education: 7.06%
- News and Media: 8.81%
- Personal: 7.04%
- Science: 7.05%
- Self-improvement: 7.93%
- Sport: 7.01%
- Technology: 7.61%

B.1 Insights from the Data

• **Wide Range of Topics**: ArPod2.0 covers diverse themes, from broad areas like ”News and Media” and Technology” to niche subjects like ”Gaming” and ”Sports.” This range caters to different audience interests.

• **Cultural and Regional Influences**: The dataset reflects various cultural and regional preferences. For instance, “Literature and Education” may be especially popular in regions with a strong emphasis on learning.

• **Language Diversity**: Including Modern Standard Arabic, regional dialects, and English expressions, the dataset appeals to a wide range of Arabic-speaking communities, enhancing its applicability.

• **Educational and Informative Content**: Topics like ”Science,” ”Self-Improvement,” and ”Literature and Education” highlight the dataset’s focus on delivering educational and enriching content.

• **Current and Relevant Themes**: The presence of ”News and Media” ensures the dataset stays relevant and updated with contemporary issues and trends.

B.2 Summary

ArPod2.0 provides a detailed and varied collection of topics, totaling 30 hours and 23 minutes across 13 categories. This diversity is influenced by factors such as content popularity, cultural context, and the mix of languages. Overall, ArPod2.0 is a valuable resource for research and exploration in spoken Arabic content.

4 Spoken Topic Identification Framework

In this section, we explain how we use audio features to identify topics in spoken Arabic content. We employ machine learning models that focus on extracting and analyzing these audio features. Our goal is to determine how well our approach can recognize topics based on audio data alone.

© Copyright HDSKD
All Rights Reserved

We evaluate the performance of our models using accuracy to show that audio information is a strong predictor for identifying spoken topics.

A. Acoustic-Spectral Cues for Spoken Topic Modeling

The core of our system is the Feature Extraction Block, which extracts key information from raw audio data. This step is crucial for identifying topics in spoken language.

The Feature Extraction Block acts as a filter that converts complex audio signals into a set of simpler, distinct features.

TABLE II :
BEST TOPOLOGY FOR EACH DESCRIPTOR

Descriptor	Parameter	Length
MFCC	• signal sampling rate • n mfcc • dct type • Chroma	40
Chroma	• sampling rate • n fft • window • n chroma	12
Tonnetz	• chroma_bins	6
Melspectrogram	• nfft • signal-sampling_rate • window	128
Contrast	• signal-sampling_rate • window • n_bands	7
Total Features	/	**193**

This process not only speeds up the analysis but also enhances its accuracy, allowing the system to capture the essential characteristics of spoken language. Essentially, this block is vital for exploring and understanding spoken topics through its detailed feature extraction process.

To determine the best settings for each classifier, we conducted a series of experiments (Lounnas 2019), (Lounnas 2022). Table II summarizes the optimal configurations for various audio descriptors like MFCC (Mel-Frequency Cepstral Coefficients), Chroma, Tonnetz, Melspectrogram, and Contrast (McFee 2015).Each descriptor has its unique set of parameters, which can be adjusted to fine-tune its feature extraction capabilities. This adjustment helps in optimizing the length and quality of the extracted features, which are crucial for accurate topic identification. Understanding these configurations allows us to choose the best settings for our spoken topic identification system. The combination of these audio features highlights the variety and complexity present in the spoken content, aiding in the accurate identification of topics.

© Copyright HDSKD
All Rights Reserved

B. Applied Machine Learning Techniques for Spoken Topic Identification

This section details how we use machine learning techniques to identify topics in spoken Arabic content. We utilize two main classifiers: Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). The specific settings for these classifiers are summarized in Table III.

C. Classifier Configurations

For the MLP classifier:

- Random State: 1
- Iterations: 300
- Solver: 'adam'
- Activation Function: 'relu'

For the SVM classifier:

- Regularization Parameter (C): 1 (low error tolerance)
- Kernel: 'RBF' (Radial Basis Function, used for handling complex patterns)
- Gamma Parameter: 'scale' (adjusts flexibility of the model) (Pedregosa 2011)

These parameters guide how each classifier learns and makes predictions. Adjusting these settings can greatly affect how well the model performs and how well it generalizes to new data.

TABLE III
BEST TOPOLOGY FOR EACH CLASSIFIER

Classifier	Parameter	Value
Multi-Layer Perceptron (MLP)	• random state • max iter	• 1 • 300
SVM	• C • kernel • gamma	• 1 • 'RBF' • 'Scale

D. Machine Learning Process

Our system for spoken topic identification operates in two main phases: training and testing, as depicted in Figure 2.

D.1 Feature Extraction

- This is the first step where we capture important acoustic and spectral features from the raw audio data.
- The extracted features form a basis for both training and testing.

D.2 Training Phase

© Copyright HDSKD
All Rights Reserved

• In this phase, we use the extracted features to train our models.
• We build an acoustic-spectral model using the features, which combines both acoustic and spectral characteristics of the audio.
• We train two classifiers: SVM and MLP, on these features to recognize patterns that are associated with different topics.

D.3 Testing Phase

• After training, we use the models to classify new audio data.
• The trained SVM and MLP models apply the learned patterns to categorize the new spoken content into its respective topics. This structured approach allows our system to effectively learn from the audio data and accurately identify topics based on the patterns present in the acoustic and spectral features.

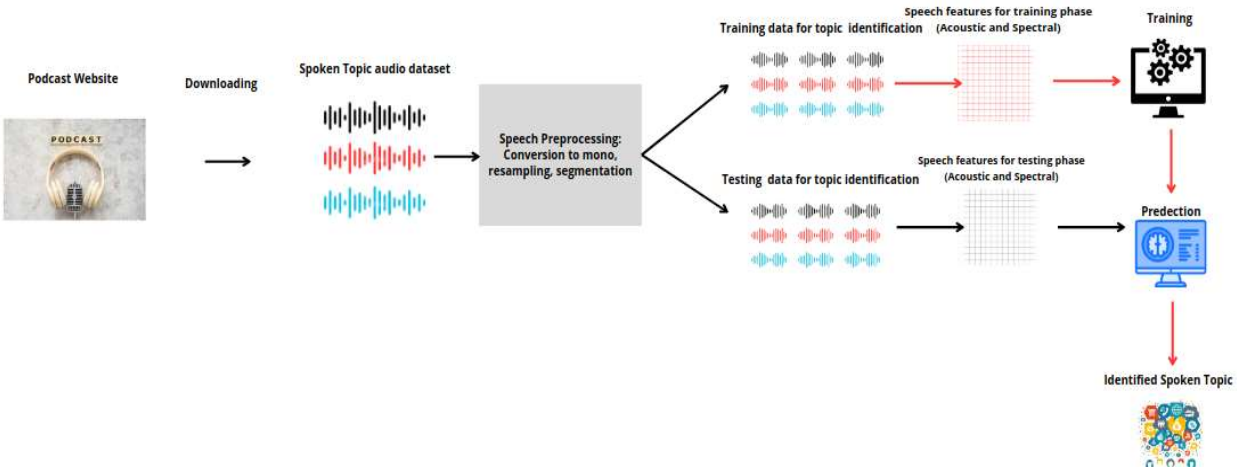

Fig. 2. Our Arabic Spoken Topic Identification Framework

5 Results and Analysis

In this section, we present the findings from our evaluation of binary and multiclass topic identification methods for spoken Arabic content. Our goal was to understand how well these methods can classify content by topic, providing insights into their strengths and limitations.

A. Binary Topic Identification Challenges

We conducted a series of 15 binary classification experiments, each comparing a pair of the six topics. This allows us to analyze how well our models distinguish between different pairs of topics. The detailed results for business-related topics are shown in Table IV.

TABLE IV
BUISNESS VERSUS ALL TOPICS

Model\Pairs	BU-LAE	BU-SE	BU-TECH	BU-GE	BU-HE
SVM	96.35	96.48	99.09	97.46	98.36
MLP	98.94	98.35	99.17	99.29	99.04

© Copyright HDSKD
All Rights Reserved

A.1 Business Topic Analysis

Both SVM and MLP models exhibited robust performance in classifying business-related topics. SVM consistently achieved high accuracy, averaging above 97.5% across all categories, with exceptional performance in the BU-TECH category (99.09%). MLP also showed strong performance, with accuracy ranging from 98% to 99% across all business topics, particularly excelling in the BU-GE category.

A.2 General Topic Analysis

In the analysis of general topics, MLP consistently outperformed SVM in all sub-topics, achieving accuracy close to or exceeding 99%. While SVM maintained high accuracy across all categories, MLP demonstrated superior adaptability to diverse and complex general topics.

A.3 Health Topic Analysis

Both SVM and MLP models demonstrated promising results in classifying health-related topics. SVM showed stable performance with accuracy ranging from 96.64% to 98.53%, while MLP surpassed SVM with accuracy scores averaging above 99.17%. MLP's higher accuracy suggests its strength in capturing complex patterns within health-related content.

A.4 Literature and Education Topic Analysis

MLP exhibited superior performance in the classification of Literature and Education topics, achieving accuracy rates of 99.29% in LAE-SE and 99.83% in LAE-TECH. While SVM also performed well, MLP's accuracy indicates its adeptness at handling technical aspects within educational literature.

A.5 Self-Improvement Topic Analysis

In the analysis of Self-Improvement topics, MLP achieved exceptional accuracy, reaching the highest accuracy rate of 99.68% in the SE-TECH sub-topic. This outstanding performance underscores MLP's effectiveness in capturing complex patterns within self-improvement content.

B. Multi-Label Topic Identification

Table V showcases the classification performance of SVM and MLP models in identifying multiple topics simultaneously. The SVM model achieved an accuracy of 93.61%, while the MLP model surpassed it with a higher accuracy of 97.62%.The results indicate that both SVM and MLP models are effective in identifying content associated with multiple topics. MLP exhibits a slight advantage, particularly in accurately classifying instances related to self-improvement and technology.

TABLE V
MULTI TOPIC IDENTIFICATION

Classifier	ALL Classes
SVM	93.61
MLP	97.62

© Copyright HDSKD
All Rights Reserved

The variation in performance between SVM and MLP models can be attributed to their respective capabilities in handling different patterns and relationships within the dataset. SVMs, proficient in scenarios with linear separability, may encounter challenges with intricate patterns, leading to fluctuations in performance across topics. Conversely, MLPs, equipped with a multi-layered structure, demonstrate greater flexibility in capturing complex, non-linear relationships, making them more adept at handling diverse topics.
These findings underscore the importance of selecting models based on the inherent characteristics of the dataset and the complexity of relationships within specific topic domains.

7. Discussion

The varied performance of machine learning models across different topics in our dataset reflects the diverse nature of the content and the inherent challenges therein. For instance, the topic ”Technology” may encompass discussions on complex subjects like artificial intelligence or cybersecurity, posing challenges in classification due to technical jargon and rapidly evolving trends. In contrast, topics like ”Health and Wellness” might present clearer boundaries with well-defined terminology, making classification relatively straightforward. However, factors such as data imbalance, where there are fewer instances of topics like ”Environmental Conservation” compared to ”Entertainment,” can lead to bias in model predictions. Moreover, topics with similar themes, such as ”Education Reform” and ”Public Policy,” may share common keywords, making it harder for models to differentiate between them accurately. Additionally, noise in the data, such as colloquial language in ”Youth Culture” discussions, can further complicate classification efforts. To address these challenges and improve overall model performance, strategies such as fine-tuning model parameters, incorporating domain-specific features, and mitigating data imbalance through resampling techniques will be essential.

8. Conclusion

In conclusion, ArPod2.0 stands as a remarkable contribution to the realm of Arabic Natural Language Processing (NLP). With its extensive corpus spanning over 30 hours, meticulously curated for Spoken Topic Identification (STI) tasks, ArPod2.0 fills a crucial gap in the availability of comprehensive Arabic speech datasets tailored specifically for STI purposes. By encompassing a diverse range of Arabic dialects alongside English, ArPod2.0 not only facilitates the development of more accurate and robust STI models but also opens avenues for cross-linguistic analysis and research. Looking ahead, the utilization of ArPod2.0 in various applications, such as enhancing speech recognition systems and enabling deeper linguistic insights, holds immense promise.

© Copyright HDSKD
All Rights Reserved

References

(Abusulaiman 2017) Abusulaiman, N. S., & Alhanjouri, M. A. (2017). Spoken Arabic news classification based on speech features. *International Journal for Research in Applied Science and Engineering Technology, 5*.

(Bost 2015) Bost, X., Senay, G., El-Bèze, M., & De Mori, R. (2015). Multiple topic identification in human/human conversations. *Computer Speech & Language, 34*(1), 18–42.

(Hazen 2008) Hazen, T. J., & Margolis, A. (2008, March). Discriminative feature weighting using MCE training for topic identification of spoken audio recordings. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4965–4968). IEEE.

(Hazen 2011a) Hazen, T. J. (2011). MCE training techniques for topic identification of spoken audio documents. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(8), 2451–2460.

(Hazen 2011b) Hazen, T. J. (2011). MCE Training Techniques for Topic Identification of Spoken Audio Documents. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(8), 2451–2460. https://doi.org/10.1109/TASL.2011.2139207

(Huijbregts 2011) Huijbregts, M., McLaren, M., & Van Leeuwen, D. (2011, May). Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4436–4439). IEEE.

(Kesiraju 2017) Kesiraju, S., Pappagari, R., Ondel, L., Burget, L., Dehak, N., Khudanpur, S., ... & Gangashetty, S. V. (2017, March). Topic identification of spoken documents using unsupervised acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5745–5749). IEEE.

(Liu 2017) Liu, C., Yang, J., Sun, M., Kesiraju, S., Rott, A., Ondel, L., ... & Khudanpur, S. (2017, March). An empirical evaluation of zero resource acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5305–5309). IEEE.

(Lounnas 2019) Lounnas, K., Abbas, M., & Lichouri, M. (2019, September). Building a speech corpus based on Arabic podcasts for language and dialect identification. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing* (pp. 54–58).

(Lounnas 2022) Lounnas, K., Abbas, M., & Lichouri, M. et al. (2022). Enhancement of spoken digits recognition for under-resourced languages: Case of Algerian and Moroccan dialects. *International Journal of Speech Technology, 25*, 443–455. https://doi.org/10.1007/s10772-022-09971-y

© Copyright HDSKD
All Rights Reserved

(May 2015) May, C., Ferraro, F., McCree, A., Wintrode, J., Garcia-Romero, D., & Van Durme, B. (2015, September). Topic identification and discovery on text and speech. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2377–2387).

(McFee 2015) McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18–25).

(Parcollet 2016) Parcollet, T., Morchid, M., Bousquet, P. M., Dufour, R., Linarès, G., & De Mori, R. (2016, December). Quaternion neural networks for spoken language understanding. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 362–368). IEEE.

(Park and Glass 2008) Park, A. S., & Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(1), 186–197. https://doi.org/10.1109/TASL.2007.909282

(Pappagari 2019) Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019, December). Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 838–844). IEEE.

(Pedregosa 2011) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830.

(Sung 2020) Sung, M. L. (2020). Unsupervised spoken term discovery on untranscribed speech. *arXiv preprint arXiv:2011.14060*.

(Wintrode 2011) Wintrode, J. (2011, May). Using latent topic features to improve binary classification of spoken documents. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5544–5547). IEEE.