INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Teaming up: making the most of diverse representations for a novel personalized speech retrieval application

Stephanie Pancoast[1]*, *Murat Akbacak*[2]

[1]Airbnb, San Francisco, CA
[2]Apple, Cupertino, CA

stephanie.pancoast@gmail.com, murat.akbacak@ieee.org

Abstract

In addition to the increasing number of publicly available multimedia documents generated and searched every day, there is also a large corpora of personalized videos, images and spoken recordings, stored on users' private devices and/or in their personal accounts in the cloud. Retrieving spoken items via voice commonly involves supervised indexing approaches such as large vocabulary speech recognition. When these items are personalized recordings, diverse and personalized content causes recognition systems to experience mis-matches mostly in vocabulary and language model components, and sometimes even in the language users use. All of these contribute to retrieval task performing very poorly. Alternatively, common audio patterns can be captured and used for exampler-based retrieval in an unsupervised fashion but this approach has its limitations as well. In this work we explore supervised, unsupervised and fusion techniques to perform the retrieval of short personalized spoken utterances. On a small collection of personal recordings, we find that when fusing word, phoneme and unsupervised frame based systems, we can improve accuracy on the top retrieved item approximately 3% above the best performing individual system. Besides demonstrating this improvement on our initial collection, we hope to attract community's interest to such novel personalized retrieval applications.

Index Terms: spoken utterance retrieval, speech indexing, unsupervised speech representation, system fusion

1. Introduction

As the world continues to move into a digital age, an increasing number of multimedia files, including spoken recordings, are created and stored. In addition to the large corpora of publicly available digital items, users also create personalized spoken recordings such as reminders or voice memos. Retrieving the files, however, can be difficult, especially when the meaningful information within the document is in the form of a continuous audio signal.

In this work we focus on extracting representations for a personalized retrieval task. Previous related work falls under two categories: supervised and unsupervised spoken information retrieval. Supervised approaches most commonly involve extracting a textual representation (either word or phoneme) from an automatic speech recognition (ASR) output. Both word and phonetic recognition output have been used for latent topic modeling for the task of document topic clustering and classification [1, 2]. The phonetic approach, although requiring less training data in comparison to the word system, was able to model the topics with a low error rate. However, in the case of language mismatch, authors [2] found performance to degrade noticeably.

In addition to supervised approaches, recent work has also explored unsupervised techniques. Authors [2] used unsupervised acoustic units as the basis for topic classification on long audio documents while other studies [3, 4] employed segmental dynamic time warping (SDTW) to perform spoken term discovery on audio lecture data. SDTW, first introduced by Park and Glass [5] using raw acoustic features, can also be applied to frame level posteriorgrams as was done by Zhang and Glass [6] to increase the method's robustness. [7] presents a nice overview of this method from robustness to speaker changes perspective.

The retrieval task addressed in this paper involves short, personalized spoken queries and items. As a personalized retrieval scenario, the speaker is the same for both the query and retrieved items. Recordings capture short descriptions in a memo format for users to remember a pointer to a personal document or item on their device. The utterances may contain frequent OOV terms, language mixing, word re-orderings, and ungrammatical sentences. In contrast to the longer spoken documents with constrained topics used in related works [1, 2, 3, 4, 5, 6], the personalized short utterances provide few opportunities to capture the repeated patterns and the topics are not constrained or well defined. Besides demonstrating the technical challenges and effective solutions for this novel voice-based retrieval application, we hope to attract community's interest to such novel personalized retrieval applications presenting interesting research problems.

We first provide an overview of different indexing representations and discuss the advantages and disadvantages of each in Section 2. In Section 3 we present the methods used to employ each representation, including a newly proposed back-off strategy to incorporate the N-best recognition outputs for the supervised approaches. Late fusion which combines the strengths of the individual representations is also discussed. Finally in Section 4 we present the results and provide a thorough analysis of the performance and errors.

2. Overview

Each individual system, whether using a supervised or unsupervised approach, has its strengths and weaknesses. The word based technique, although carrying inherent semantic meaning, requires significant training data. Further, recognition fails when there is a language mismatch and struggles in the presence of out-of-vocabulary (OOV) terms. In one example two

* This work was done as part of a summer internship when both authors were at Microsoft. First author was a PhD student at Stanford University at the time and this work was part of her PhD thesis.

http://dx.doi.org/10.21437/Interspeech.2016-1589

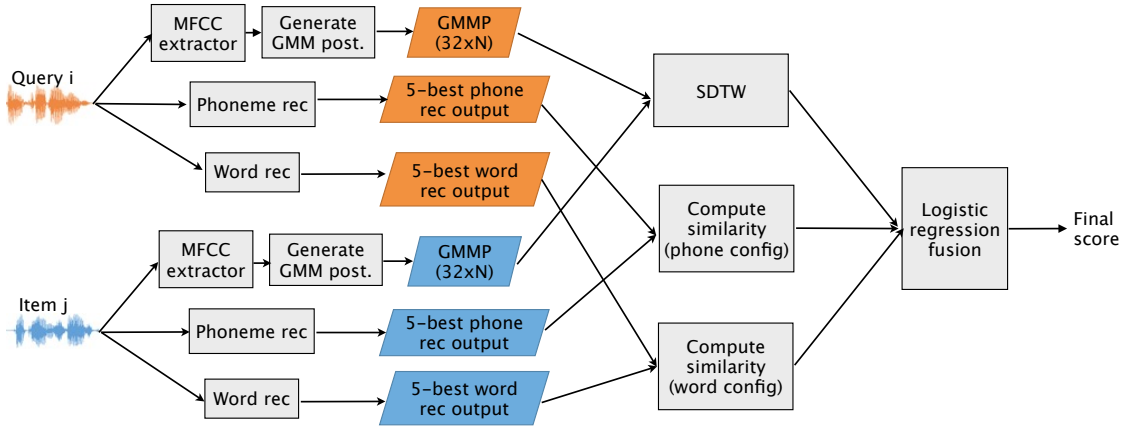

Figure 1: Diagram of individual systems and fusion for short spoken utterance retrieval. Representations (32-dimensional GMM posteriorgrams for each of the N frames in an utterance, phoneme recognition output, word recognition output) are first extracted for the query and item. A similarity score is calculated for each representation individually, and then the scores are fused using logistic regression to produce the final similarity score.

utterances are nearly identical:"Noha Alon One Note page" and "Noah's One Note Page." Yet the pair only has a single overlapping word between the 5-best recognition outputs. The name "Noha" is an OOV and, by having the last name in only one of the utterances, the recognition outputs can no longer be matched. Phoneme level approaches are more flexible but as presented by Hazen et. al. [2] still struggle in language mismatch scenarios. Finally, the unsupervised frame approaches such as SDTW have the advantage of not requiring any supervised training. As a tradeoff, matches tend to be highly sensitive to the presence of noise and are more likely to be hallucinated than the supervised based techniques.

No individual representation is perfect, yet the strengths and weaknesses of each complement one another. For this reason, we also explore fusion. Recent research has combined different systems. Typically the combination involves either ASR and phone-based systems [8], or phone-based and unsupervised systems [9].

Illustrated in Figure 1 is an overview of retrieval using the individual representations and late fusion. Two spoken utterances, query i and item j, are entered into the three individual representation systems: unsupervised frame (top), phone (middle), and word (bottom). A similarity score for each level is computed separately, and then combined using logistic regression. These steps are all described in detail in the remainder of this paper.

3. Method

Spoken utterance retrieval, illustrated in Figure 1, involves two main steps: generating the representations and calculating the similarity score for all query-item pairs. In this section we describe these retrieval components.

3.1. Representation Extraction

Before comparing utterances, the initial representations first need to be extracted. For the phoneme and word level this is the 5-best recognition output from a phonetic recognizer (for the phonetic approach) and a large vocabulary continuous speech recognition system (for the word approach).

The unsupervised frame approach represents the speech document as a series of Gaussian posteriorgrams vectors. Gaussian mixture models (GMMs) are a very common frame-level representation for speech signals and have long been used in speech-related tasks. Zang et. al. [6] found the posteriorgram representation to outperform MFCCs when applying SDTW so we also use this approach. The GMM posteriorgram is trained using a held-out tuning set on the first 12 MFCCs plus the log energy extracted from 10ms-long speech frames (discarding non-speech frames) and has 32 components and diagonal covariance.

3.2. Representation Matching and Retrieval

The representations are not easily comparable in their original forms and first need to be transformed. This step varies by system, as does the similarity score computation. The system-specific transformation and matching are described in more detail in the remainder of this section.

3.2.1. Word and Phoneme Level

The word- and phone- level representations are similar in that both consist of a 5-best output from a recognition system along with word or phone confidences. Although parameters vary (e.g. N-gram size), the approaches and enhancements explored for each are the same.

To compute the similarity of the 5-best recognition output for a given utterance, all N-grams are assigned a confidence score based on the word- or phone- level confidences. As described in Equation 1, the similarity score is calculated by finding the common N-grams in the pair, and summing over the confidences for each N-gram. Length normalization is also applied for the respective utterance.

$$score_N(q,d) = \frac{\sum_{i:d_i \in M} conf(d,i)}{\|d\|} + \frac{\sum_{j:q_j \in M} conf(q,j)}{\|q\|} \quad (1)$$

Here q and d denote the query and document, M is the set of overlapping N-grams length N, $conf(d,i)$ is the confidence

of N-gram i in document d, and $\|d\|$, $\|q\|$ are the total number of N-grams generated from the 5-best output in item d and query q respectively.

A common practice in natural language processing is to combine N-grams of various sizes to represent the document. Equation 2 calculates a single score when considering different N-gram sizes, weighting each N-gram by its length to consider longer matches more strongly than shorter ones.

$$score(d, q) = \sum_{n} n \bullet score_n \tag{2}$$

We explored two weighting schemes to be applied to the word and phone level matching: backoff and inverse document frequency (IDF) weighting. When an N-gram appears in all five outputs of the 5-best recognition counting it five times towards a match score yields a stronger weight on that single N-gram than desired. We apply backoff to discount the contribution of an N-gram for each appearance in the 5-best output. If $conf(d, j)$ is an N-gram's score, for the i^{th} appearance of that N-gram in the recognition of document d, the score is updated as $conf(d, j) \leftarrow conf(d, j) + \frac{c}{2^i}$ where c is the original confidence of the N-gram.

IDF weighting is a common technique applied for information retrieval tasks and is also used for spoken documents [1]. If an N-gram that is present in most documents (high document frequency) appears in the query-item pair, it is not as strong an indication of a match as a pair that contains a unique N-gram. IDF weighting divides the original confidence by the document frequency when computing the score.

3.2.2. Unsupervised Frame-Level Alignment

Segmental dynamical time warping (SDTW), first used by Park and Glass [5], finds the best matching pattern between two speech-containing audio segments. The SDTW algorithm first generates a similarity matrix (also referred to as a *dotplot* [4]). Each element (i, j) in the matrix is equal to the distance between frame i and frame j, according to some distance metric appropriate to the original acoustic feature.

SDTW relies on the assumption that common speech patterns are acoustically similar to each other in the original feature space. The algorithm finds the path of length greater than L within a sub-band of width W that has the minimum distortion in the similarity matrix. The parameters L and W are predetermined and ensure the discovered pattern is long enough to be meaningful as well as temporally-constrained. After running SDTW on a similarity matrix, a set of alignments and their distortion scores remain, each representing a hypothesized match.

To generate the final similarity score for the unsupervised frame level, we take the lowest-distortion alignment for a single pair. Note that this score has the opposite relation of the word- and phone- level in that a lower value indicates a stronger match.

3.3. Score Fusion

As mentioned previously in Section 2, each individual representation has advantages and disadvantages. By combining, or fusing, the the similarity scores of the three approaches we hope to capitalize on the strengths of the word-, phone- and unsupervised frame- systems to generate a more robust score.

A common method for score fusing applies logistic regression, as calculated by Equation 3, to the subsystem scores to generate a single similarity measure.

$$F(\mathbf{x}) = \mathbf{e}^{\beta \bullet \mathbf{x}} \tag{3}$$

Here the vector $\mathbf{x}$ contains the representation similarity scores and the vector β contains the coefficients. Logistic regression requires held-out data to compute these coefficients.

4. Experimental Results and Analysis

For personalized spoken utterance retrieval there exists a database of digital items linked with a short recording, and upon receiving a spoken utterance as the query, the items are ranked according to a similarity score so that the top-K can be returned. We measure performance as accuracy of the top-1 retrieved item. Each spoken item and query is converted into a representation as previously mentioned in Section 3.1. For each query in the dataset, the similarity score is calculated for every document according to Equations 1 and 2. If the highest scoring item is a match, we consider it a success, otherwise it is a failure.

4.1. Data

Our data consists of 72 spoken items and 247 queries, collected across three subjects. The data used in this work is the transcribed subset of a larger dataset. In order to provide a thorough analysis, transcriptions were valuable and so the experiments were restricted to these queries and items. During data collection, the subject was shown a personalized digital document (e.g. an image) and was asked to create a brief spoken description up to 5 seconds of the displayed document. Over the following few days, the subject entered multiple spoken queries in an attempt to retrieve the documents. The average duration is 2.56 seconds for the queries and 2.60 seconds for the database items. Since we are interested in personalized utterance retrieval, the only items considered for a query are those from the same speaker.

4.2. Representations

Results for all three representations are presented in Table 1 along with the fusion results and breakdown by speaker. For the word based system, unigrams through trigrams were tried and it was found that the combination of unigram, bigrams and trigrams showed the best result, especially when applying both backoff and IDF weighting. This configuration resulted in a top-1 retrieval accuracy of 86.38%. Because the duration of phonemes is shorter than words, longer N-grams were explored for this subsystem. Different combinations of N-grams with the minimum ranging from one to five and the maximum from three to ten were explored, with a combination of N-grams length three to five performing the best. Like with the word system, backoff and IDF weighting both improved performance yielding an accuracy of 94.37%. The unsupervised frame level system using parameter values from [5] (L=50 and W=5) yielded 81.46% accuracy.

The per-speaker performance and error analysis exemplifies the strengths and weaknesses of each representation. At the frame level, errors are predominately due to the confusable items where the majority of the sentence overlaps and the ability to discern between the two depends heavily on identifying a small portion of the utterance. For example, two of *Speaker A*'s items ("Nohas One Note" and "Yuzongs One Note") differ by only one term, which also happens to be an OOV for both. The queries for *Speaker B*, on the other hand, are nearly al-

System	*Sp. A*	*Sp. B*	*Sp. C*	Overall
Frame	58.06	95.51	75.29	81.46
Phone	81.25	97.92	95.29	94.37
Word	62.50	84.46	87.06	86.38
Late fusion	90.32	100.00	96.47	97.07

Table 1: Top-1 retrieval accuracy for single representation (unsupervised frame, phoneme, and word) and late fusion experiments. Results are also broken down by individual speaker (*Sp*).

System	*Sp. A*	*Sp. B*	*Sp. C*
Frame	-2.48	-4.22	-2.66
Phone	1.45	5.08	1.40
Word	-0.47	3.04	0.86

Table 2: Logistic regression coefficient averages for late fusion 5-fold cross validation experiments. Since the coefficients are trained separately for each speaker (*Sp.*) we present the coefficient means individually.

ways verbatim of the matching item. *Speaker C*, who has more confusable items than C but less than *A* shows an accuracy in between the other two subjects.

At the other end of the spectrum, the word level is heavily influenced by language mismatch and word re-ordering in the presence of OOVs. Over 30% of *Speaker A*'s items are in a foreign language (Turkish) in contrast to the 10% for *Speaker B* (Hebrew). *Speaker C* mixes in some other language terms (Spanish) within a single utterance, but has no fully non-English recordings and as a result has the highest retrieval performance at the word level. It is at the phoneme level, a compromise between the two extreme representations, that the top-1 retrieval accuracies are the most similar. As is clear from Table 1 and discussed in the next section, the advantages and disadvantages of each representation can be capitalized on by late fusion and outperform the individual representations for all three speakers.

4.3. Late Fusion

Because of the limited data available, we use 5-fold cross validation for the late fusion. For each fold and each speaker, the logistic regression coefficients are trained on $\frac{4}{5}$ of the data and the final late fusion similarity scores are generated on the remaining $\frac{1}{5}$ using this model. Results from score fusion are presented in the last row of Table 1. In comparison with the single representation results we can see the clear improvement gained by applying late fusion, increasing the overall retrieval accuracy from 94.37% for the best single representation system to 97.07% for late fusion. To further confirm that all three representations all valuable, experiments were run leaving one out and found that omitting any one subsystem decreases late fusion performance.

In logistic regression, the weights can provide insight into the individual contribution and stability of the different representations. The statistics for the word-, phone- and unsupervised frame-level system scores for the three speakers are presented in Table 2. These further confirm observations from the top-1 retrieval performance of the individual representations. The frame-level coefficients are negative because, as previously mentioned, for this subsystem lower scores are indicative of stronger matches. We observe the word-based system is the least reliable of the three, especially for *Speaker A* who has a significant portion of non-English utterances. Also, as evident by the individual performances presented in Table 1, the phoneme representation, a compromise between the two extremes, performs the best for all three speakers.

5. Conclusions and Future Work

In this work we presented results for a novel personalized speech retrieval application where users create memo-style recordings to remember a pointer to a personal document or item on their device. Proposed approach employed word, phoneme and unsupervised frame representations as subsystems. We demonstrated the effectiveness of late system fusion of these different representations on a small but diverse speech recordings where subjects presented different use cases (e.g., some subjects introduced word re-orderings, some subjects introduced foreign words or languages, etc.). Late fusion yielded more than 3% increase in top-1 retrieval accuracy over the best-performing representation. Beyond accuracy metrics, we included a thorough analysis of the performance of the proposed techniques. Besides demonstrating technical challenges and effective solutions proposed for this novel personalized speech retrieval application, we hope to attract community's interest to such personalized speech retrieval applications where efforts in different areas of speech processing and recognition, as well as efforts in information retrieval and machine learning fields can team up to tackle the novel technical problems introduced by the nature of this application.

As future work, instead of late fusion which only considers the final similarity score for each representation and therefore discards the temporal agreement, we would like to consider building mid-level fusion approach where the final similarity score used for retrieval performs the calculation at an earlier stage which can account for the temporal agreement (e.g. does the match at the word level agree with the match at the phone level). Another direction we would like to work on is borrowing adaptation techniques from web search, which will be important especially when personalized speech retrieval applications are widely deployed and users provide click information as they find items they are searching for. This is an important aspect as users will create personal recordings overtime and there will be opportunities for individual systems as well as the fusion component to learn from the click logs every time users try to retrieve an item and interact with the system.

6. Acknowledgments

We would like to thank several ex-colleagues at Microsoft speech team for their helpful feedback, and especially Noha Alon for building the mobile app we used for data collection. This material is also based upon work supported by the National Science Foundation under Grant No. DGE-1147470.

7. References

[1] T. Hazen, "Direct and latent modeling techniques for computing spoken document similarity," in *Spoken Language Technology Workshop (SLT)*, 2010.

[2] T. Hazen, M. Siu, H. Gish, and S. Lowe, "Topic modeling for spoken documents using only phonetic information," in *Automatic speech recognition and understanding (ASRU)*, 2011.

[3] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

[4] A. Jansen, K. Church, and H. Hynek, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[5] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[6] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on gaussian posterior grams," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009.

[7] P. F. Schulam and M. Akbacak, "Diagnostic techniques for spoken keyword discovery," in *Interspeech*, 2014.

[8] S. Jin and T. Sikora, "Combining confusion networks with probabilistic phone matching for open-vocabulary keyword spotting in spontaneous speech signal," in *European Signal Processing Conference*, 2009.

[9] H. Wang, T. Lee, C. Leung, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.