DCU at the NTCIR-11 SpokenQuery&Doc Task

David N. Racca
CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University
Ireland
dracca@computing.dcu.ie

Gareth J.F. Jones
CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University
Ireland
gjones@computing.dcu.ie

ABSTRACT

We describe DCU's participation in the NTCIR-11 SpokenQuery&Document task. We participated in the spoken-query spoken content retrieval (SQ-SCR) subtask by using the slide group segments as basic indexing and retrieval units. Our approach integrates normalised prosodic features into a standard BM25 weighting function to increase weights for terms that are prominent in speech. Text queries and relevance assessment data from the NTCIR-10 SpokenDoc-2 passage retrieval task were used to train the prosodic-based models. Evaluation results indicate that our prosodic-based retrieval models do not provide significant improvements over a text-based BM25 model, but suggest that they can be useful for certain queries.

Team Name

CNGL

Subtasks

SQ-SCR over slide group segments (Japanese)

Keywords

spoken content retrieval, weighting functions, prosody

1. INTRODUCTION

The NTCIR-11 SpokenQuery&Doc Task [2] provides a common evaluation framework for researchers interested in tasks of spoken term detection (STD) and spoken content retrieval (SCR) over collections of informally structured spoken content. The organisers provided retrieval subtasks which require SCR systems to search for either relevant content or term mentions within a collection of lecture recordings in Japanese language.

In the spoken-query spoken content retrieval (SQ-SCR) subtask, spontaneous spoken queries were provided as input to retrieval systems whose goal is to find relevant predefined speech segments within a collection of lecture recordings. Participants were required to choose between two predefined retrieval units:

1. Slide group segments: speech segments with time boundaries given by the start and end points of a group of topically coherent contiguous slides in the lectures.

2. Inter pausal units (IPUs): speech segments obtained by splitting the speech data between silences longer than a pre-computed threshold.

We participated in the SQ-SCR subtask and chose the slide group segments as the basic indexing and retrieval units for our experiments. More information about the tasks, dataset, queries, transcripts, and evaluation metrics can be found in the overview paper [2].

In a typical SCR system, spoken documents are initially converted into text transcripts by means of large vocabulary continuous speech recognition (LVCSR) systems. Normally, these transcripts do not include other information than sequences of hypothesized words in the form of N-best lists, confusion networks, or lattices, with some additional timing and confidence information. This representation of speech is an oversimplification of spoken language which is well known to encode richer information that goes beyond the lexical level of words and syllables. In particular, one information source that is not represented in LVCSR transcripts, and that has not been exploited in previous work on SCR is prosodic information, which is characterised by variations in pitch, duration, and loudness of the spoken units across time. Prosody is used in human communication for a wide range of purposes, including, among others, disambiguation of meaning in relative clauses, making contrastive emphasis or focus, and structuring information into phrases [18].

In this work, we build upon techniques proposed in [7] and [3] to explore the utility of prosodic features in the task of retrieving relevant speech segments. The approach integrates prosodic features into a BM25 weighting function [13] to give higher weight to terms that are prosodically prominent in the spoken content, thus promoting the rank of documents that contain a high number of prominent terms. This paper describes our approach and reports experimental results in the SQ-SCR subtask by using slide group segments as search units.

This paper is structured as follows. Section 2 presents relevant previous and related work, while Section 3 explains in detail our approach and different runs submitted for evaluation. Section 4 presents and discusses the evaluation results in terms of the official evaluation metric of the task. Finally, Section 5 presents our conclusions and suggestions for further investigations.

2. BACKGROUND AND PREVIOUS WORK

This section begins with a brief description of general TF-IDF weighting schemes in SCR in Section 2.1. At the same time, we introduce some notation that will be used throughout the paper. Section 2.2 gives basic notions about prosody and prosodic prominence while Section 2.3 describes previous research that investigated the relationship between

prosodic prominence and TF-IDF weights. Lastly, Section 2.4 presents related and previous research carried on tasks that are relevant to SCR, where researchers have tried to integrate prosodic information into TF-IDF weighting schemes.

2.1 TF-IDF Weighting Schemes in SCR

Given a collection of spoken documents segmented into a collection C of topically consistent segments $\mathbf{s}_1, \ldots, \mathbf{s}_N$, the task of an SCR system can be stated as to produce a ranked list of segments in C that are relevant to a query $\mathbf{q}$ provided by a user. In order to achieve this, an SCR system typically implements an information retrieval (IR) component that computes relevance scores for each segment with respect to the query. This is, given a query $\mathbf{q}$, the IR component has to rank the segments in C according to the value returned by the relevance score function:

$$\mathrm{rel}(\mathbf{q}, \mathbf{s}_j) \qquad j = 1, \ldots, N \tag{1}$$

A number of retrieval frameworks have been developed to define the relevance score function. Three standard IR frameworks are the vector space model (VSM) [15], the probabilistic approach [17], and the language model [11].

What most standard retrieval frameworks have in common, except for some N-grams based language models, is that they adopt the term independence assumption. This assumption states that terms occur independently from each other in the collection of segments. The independence assumption permits us to compute $\mathrm{rel}(\mathbf{q}, \mathbf{s}_j)$ as the combination of the individual contributions from the terms that appear in both $\mathbf{q}$ and $\mathbf{s}_j$. Therefore, given a list of the terms that occur in C, such as $t_1, \ldots, t_M$, we can think of any segment $\mathbf{s_j}$ in C as an M-dimentional vector where the i-th dimension is the contribution that the term t_i would make to the computation of the relevance score between $\mathbf{s_j}$ and $\mathbf{q}$.

For the query $\mathbf{q}$ and the segment $\mathbf{s}_j$ in C, the relevance score function can be written as:

$$\mathrm{rel}(\mathbf{q}, \mathbf{s}_j) = \sum_{i}^{M} \mathrm{w}(i, j) \tag{2}$$

where $\mathrm{w}(i, j)$ is the weight of term i in segment $\mathbf{s}_j$.

Ideally, term weights should characterise how well a term represents the topic of the segment in which occurs, and how well it discriminates the content of the segment from the content of other segments in the collection. Thus, the computation of the weight for the term i in $\mathbf{s}_j$, usually involves the product between the following two functions:

- $\mathrm{tf}(i, j)$, which depends on the term frequency of the term in $\mathbf{s}_j$
- $\mathrm{idf}(i, C)$, which depends on the inverse document frequency of the term in the collection.

In the rest of the paper, we will use TF-IDF score to refer to term weights based on the product:

$$\mathrm{w}(i, j) = \mathrm{tf}(i, j) * \mathrm{idf}(i, C) \tag{3}$$

independently on the concrete definition of $\mathrm{tf}(i, j)$ and $\mathrm{idf}(i, C)$.

2.2 Prosodic Prominence

In linguistics, prosody is commonly referred to as the "suprasegmental" characteristics of speech [9]. These are features that cannot be characterised as discrete speech units (segments), such as vowels or consonants, but that rather occur simultaneously with them, spanning to multiple segments, and describing their rhythmical properties. Prosodic features are not, in general, absolute characteristics of a single segment but they rather describe relative differences between segments by the way they are pronounced. For instance, vowel length or the relative duration of a syllable is considered a prosodic feature because it varies depending on the speaking rate of the speaker in a particular context.

Prosodic variation is known to be realised by varying the pitch, length, and loudness by which speech units are pronounced. The acoustic correlates of these features, which can be automatically extracted from the speech signal, are respectively the fundamental frequency (F0), duration, and signal amplitude.

A speech unit is said to be prosodically prominent when it stands out from its surrounding context by means of its prosodic characteristics. Speakers can make a word or syllable more prominent than others in order to perform different communicative functions in spoken language [8, 18]. For instance, prosodic prominence may be used by a speaker to highlight words that include new or previously given information. The general trend is that words carrying new information are more likely to be accented, while words that present old information are more likely to be deaccented. Although researchers have pointed out that there are many exceptions to this trend [8, 18], the idea that prosodic information might help to signal informative words is appealing for tasks like SCR, where commonly only simple text-based statistics are used to compute the term weights.

2.3 Relationship Between Prosodic Prominence and Term Weights

Crestani and Silipo previously investigated the potential of prosodic information in speech search by studying the relationship between stressed syllables and TF-IDF scores of terms in a corpus of american English monologues [4, 16]. In their work, linguists manually labelled every spoken syllable in the corpus as either containing a primary stress, an intermediate stress or the abscence of it, these events were given values of 1, 0.5, and 0, respectively. They then defined the stress of a word occurrence as the sum of the stress values from its syllables. The overall stress score of a word in a monologue was then defined as the average stress across all its occurrences in the monologue. Stress scores of words were then compared against their TF-IDF scores which were computed by using a BM25 weighting function and by considering monologues as the basic unit to be the indexed and retrieved.

From this comparison, the authors found that, in general, words with high (low) TF-IDF scores also tend to have high (low) stress scores. Although they could not find enough evidence to support the analogous case, their work suggests that prosodic features may have potential to identify acoustic "keywords" in the spoken content, and that this information could potentially be exploited by a SCR system to better index spoken documents.

2.4 Prosodic Prominence in Term Weighting

In [3], Chen et al. describe an approach to spoken document retrieval (SDR) that takes into account the signal amplitude and duration of words to compute relevance scores for spoken documents. In these experiments, they used a

vector space model (VSM) with cosine similarity in order to compute the relevance scores. Assuming that words containing high informative content might be uttered louder and lenghtened more than non-informative words, they increased the weight of words that were pronounced with a high average signal magnitude and large duration. The authors evaluated this method by performing retrieval experiments over a dataset of Mandarin broadcast news, and reported some minor improvements in retrieval efectiveness due to the inclusion of the prosodic features into the calculation of term weights.

More recently, Guinaudeau and Hirschberg [7] experimented with an approach that uses acoustic correlates of pitch and loudness for computing topic similarity between vector representations of spoken documents in a topic tracking task. Similarly to [3], Guinaudeau used a VSM to represent the text transcripts obtained from an automatic speech recogniser (ASR) as term incidence vectors, and considered the cosine similarity as topic similarity between transcripts.

In Guinaudeau's approach, the assumption was that terms that are more characteristic of the topic of a document might be produced "with greater emphasis, in an expanded pitch range, or with greater intensity". Therefore, they hypothesised that an acoustic score that captures how prominent a term is from its surroundings might be used in combination with TF-IDF scores to compute incidence vectors that better represent the topic of a document. In order to test this hypothesis, the authors experimented with two methods to obtain the acoustic score of a term in a document.

In their first method, they extracted the fundamental frequency (F0) and root-mean-squared (RMS) energy for every 10 ms of speech data. Subsequently, they normalised these values according to each speaker based on the output of a speaker diarisation system. This was performed by replacing each F0 and energy value by its z-score ($(v-\mu)/\sigma$), which considered the mean and standard deviation calculated across all the F0 and energy values corresponding to a single speaker.

In the next processing step, they processed the text transcripts with a stemming algorithm and aligned the normalised values of F0 and energy with each occurrence of a stem in the transcripts. As a result of this process, multiple values of F0 and energy were associated with every stem occurrence. To make the following steps clearer for the reader, let $\mathbf{f0}^k_{i,j}$ and $\mathbf{e}^k_{i,j}$ be vectors containing, respectively, the normalised F0 and energy values associated with the k-th occurrence of the i-th stem in the j-th document, and let max, min, mean, and std be functions that take a vector as an input and return, respectively, the maximum, minimum, mean, and standard deviation from all the elements of the vector. A single value of pitch and energy was given to each stem occurrence by considering the max, min, mean, and std functions computed over $\mathbf{f0}^k_{i,j}$ and $\mathbf{e}^k_{i,j}$. This resulted in four pitch scores and four energy scores for every stem occurrence. They then defined the acoustic score of the k-th occurrence of the i-th stem in document j, denoted by $\mathrm{ac}^{f,k}_{i,j}$, as the product of its pitch and energy scores as given by the function f. This resulted in four different acoustic scores, accordingly:

$$\mathrm{ac}^{f,k}_{i,j} = f(\mathbf{f0}^k_{i,j}) * f(\mathbf{e}^k_{i,j}) \quad f \in \{\max, \min, \mathrm{mean}, \mathrm{std}\} \quad (4)$$

Finally, they considered two definitions for the overall acoustic score of a term i in a document j: the average acoustic score across term occurrences,

$$\mathrm{ac}(i,j) = \frac{1}{\mathrm{tf}_{i,j}} \sum_{k}^{\mathrm{tf}_{i,j}} \mathrm{ac}^{f,k}_{i,j} \quad (5)$$

and the maximum acoustic score,

$$\mathrm{ac}(i,j) = \max\{\mathrm{ac}^{f,k}_{i,j} : k = 1, \ldots, \mathrm{tf}_{i,j}\} \quad (6)$$

Guinaudeau's second method to compute acoustic scores, was based on prominence scores calculated by the AuToBI [14] tool for every stem occurrence in an utterance, to obtain $\mathrm{ac}^k_{i,j}$. Stem duplicates in this case were handled as described previously, and either Equation 5 or 6 were used to compute the overall acoustic score of a term in a document.

In order to integrate acoustic scores into the computation of the term weights in the VSM, the authors used a harmonic mean, where TF-IDF and acoustic scores were weighted by the parameters θ_{ir} and θ_{ac} respectively in order to control their individual effects over the final score:

$$\mathrm{w}(i,j) = \frac{\theta_{ir} * \mathrm{tf}(i,j) * \mathrm{idf}(i,C) + \theta_{ac} * \mathrm{ac}(i,j)}{\theta_{ir} + \theta_{ac}} \quad (7)$$

Using Equation 7 as the weighting function for the VSM, Guinaudeau reported some improvements in terms of F1-score when using the first approach, this is, the one that does not use AuToBI, to compute the acoustic scores. Interestingly, the best results were obtained for $\theta_{ac} = \theta_{ir} = 1$, with $f = \max$ and Equation 6 to define the final acoustic score for terms.

In recent work [12], we experimented with Guinaudeau's approach [7] and implemented an SCR system that integrates normalised pitch, loudness, and duration in the computation of term weights. In a similar way, the prosodic features were used to increase the weights of prominent terms by using the harmonic mean from Equation 7. We explored different combinations of prosodic features to define the acoustic score $\mathrm{ac}(i,j)$, and evaluated these retrieval models in an ad-hoc spoken content retrieval task at the MediaEval 2014 Search and Hyperlinking benchmark [5]. Evaluation results over the test set showed that the prosodic-based systems did not offer any improvements over a text-based SCR system. However, we noted that the queries and relevance assessments used to optimise the parameters θ_{ir} and θ_{ac} in the training set differed from the type of queries and relevance assessment that were used to evaluate the models in the test set.

3. METHODOLOGY

Firstly, we processed the manual and automatic transcripts of the spoken queries and lectures to obtain tokenised text that is suitable to be indexed by an SCR system. We explain these pre-processing steps in Section 3.1. Secondly, we partitioned the transcripts of lectures based on the slide group information provided by the organisers. As a result, we obtained the collection C of segments $\mathbf{s}_1, \ldots, \mathbf{s}_N$ that we used in our retrieval experiments. Thirdly, we extracted loudness and F0 features from the audio data, and then normalised and aligned them with the words found in the segments. This is explained in more detail in Section 3.2. Lastly, we indexed the collection of segments, storing the prosodic features in the index, and performed retrieval by using a modification of the BM25 weighting scheme that incorporates acoustic scores computed from the prosodic fea-

tures of each term. Section 3.3 describes in detail how we did this while Section 3.4 explains how we selected the models submitted for evaluation.

3.1 Pre-processing of Spoken Query and Lecture Transcripts

The organisers of the SpokenQuery&Doc task provided manual and automatic speech recognition (ASR) transcripts for the spoken queries and the lecture recordings. The automatic transcripts were produced by the Julius[1] LVCSR system under different training conditions of language and acoustic models. ASR transcripts included up to 10-best recognition hypothesise for each transcribed IPU, plus confidence scores for each recognised word, and time stamps for individual words obtained from forced alignment. Recognised text was post-processed by the organisers with the ChaSen[2] morphological analyser version 2.4.4 in order to tokenise text into words and to provide base forms and part-of-speech information.

In this work, we used only the 1-best hypothesis from the word-based ASR transcripts. We extracted the base form of the recognised words, along with the force alignment information to create a linear text transcript with timing information for each word, including starting times and durations. When ChaSen was unable to produce a base form for a word, we extracted its conjugated form instead, this was done in order not to lose any possible term occurrence that may have been spoken in the lectures. As only ASR transcripts trained under the *match* and *unmatchAMLM* conditions contained timestamps for recognised words, we limited our runs to these two, omitting transcripts obtained under the *unmatchLM* condition.

We also experimented with the manual transcripts provided by the organisers. As these solely contained non-processed Japanese text, we used ChaSen version 2.4.4 to tokenise the text, though, using a different grammar, (Ipadic 2.7.0), than the grammar used by the organisers (UniDic 1.3.9) to process the ASR transcripts. Next, we used the base form of the words, when available, to create a linear text transcript for each lecture. Similarly to what we did with the ASR transcripts, the conjugated form of a word was used when ChaSen could not output its base form. In order to get the starting time and duration for each word, we first used ChaSen to obtain the pronunciation form of the word in Katakana characters, which we then translated into its phonemic representation. For instance, the sequence "m a z u i" was used as the phonemic representation for the Katakana sequence "マズイ" corresponding to the word "まずい". We then performed forced alignment with Julius and the *Julius 4 Segmentation Kit*[3] version 1.0. In this step, we obtained timestamps for each word in an IPU by feeding the julius segmentation script with the phonemic translations and the WAV file of the IPU, plus the triphone acoustic model trained under the *match* condition that was provided by the organisers.

Various types of transcripts for the spoken queries were made available for participants. We processed manual, *match* and *unmatchAMLM* ASR transcripts of queries in the same way as we did with the transcripts of lectures.

As an additional pre-processing step for manual transcripts of queries and lectures, we discarded some annotation labels that were present. In particular, annotations with label codes from the set: { H, Q, FV, 息, 笑, 泣, 咳, D, D2, ?, F, M, O, K, 笑, 泣, 咳, あくび, L, s, VAD, 雑音, H } were discarded from the manual transcripts, since we did not want them to be included in the index. As we found that the annotations in the text affected ChaSen's output, we removed them prior to performing the morphological parsing. Annotations with label codes A and W marked usage of alphabetic characters (borrowed words) and incorrect pronunciations respectively. A annotations, e.g. (A ティーエフアイディーエフ;tf-idf), specified two forms for a word, its alphabetic form and its pronunciation form in Katakana characters, while W annotations, e.g. (W エーキュー;要求), specified the mispronounced word and its correct pronunciation. Although we set-up ChaSen to omit these annotations too, we considered the alphabetic forms of words and correct pronunciations as indexing terms. Furthermore, when performing forced alignment to obtain timestamps for these special words, we used their pronunciation form included in their A or W annotations.

As the last processing step prior to indexing, we processed the spoken query and lecture transcripts with a script that converted simple-width alphabetic characters into their full-width Unicode equivalent. This was done in order not to miss trivial matchings between words containing alphabetic characters.

3.2 Processing of Prosodic Features

We extracted acoustic correlates of pitch and loudness from the audio files of each IPU by using the Munich Versatile and Fast Open-Source Audio Feature Extractor (OpenSMILE)[4] [6]. Initially, the audio signal was framed into overlapping windows of 50 ms length and 40 ms of overlap. To compute loudness, we first computed the RMS energy for each frame and then calculated its simplified intensity (narrow band approximation) by means of the OpenSMILE component *cIntensity*. Fundamental frequency (F0) was extracted for each frame by first applying the *cTransformFFT*, *cFFTmagphase*, and *cSpecScale* components to obtain octave-scaled magnitudes and phase values, and then by using the *cPitchACF* component to produce the F0 contour and probability of voicing. In the F0 contour, regions with probability of voicing below 0.55 were considered voiceless and were assigned a F0 of 0. The resulting contours of loudness and F0 were smoothed using the *cContourSmoother* component with a moving average window of size 3. At the end of this process, we obtained values of loudness and F0 for every 10 milliseconds of speech for each IPU.

The smoothed contours of loudness and F0 were subsequently aligned following a similar procedure to the one presented in Section 2.4. Following the notation convention that we introduced in Section 2.4, $\mathbf{l}^k_{i,j}$, and $\mathbf{f0}^k_{i,j}$ will denote the vector of normalised loudness and F0 values respectively associated with the k-th occurrence of i-th word in the j-th segment. These normalised vectors were obtained by using range normalisation, as follows:

$$\mathbf{l}^k_{i,j} = \frac{\mathbf{l} - \mathrm{L}_{min}}{\mathrm{L}_{max} - \mathrm{L}_{min}} \qquad \mathbf{f0}^k_{i,j} = \frac{\mathbf{f0} - \mathrm{F0}_{min}}{\mathrm{F0}_{max} - \mathrm{F0}_{min}} \tag{8}$$

where L_{max}, $\mathrm{F0}_{max}$, L_{min}, $\mathrm{F0}_{min}$ are, respectively, the ab-

[1] http://julius.sourceforge.jp/
[2] http://chasen-legacy.sourceforge.jp
[3] http://sourceforge.jp/projects/julius/downloads/32570/julius4-segmentation-kit-v1.0.tar.gz
[4] http://opensmile.sourceforge.net

solute maximum and minimum values of loudness and F0 in the lecture where the j-th segment belongs to. Note that in range normalisation, features were replaced by values in the range between 0 and 1.

In addition, we also considered the duration of words as another useful feature to be considered when computing the acoustic scores. We use $\mathrm{d}^k_{i,j}$ to denote the absolute duration of the k-th occurrence of term i in segment j.

In order to associate a single score of loudness and F0 to a specific word occurrence in a segment, we applied the max and min functions over the vectors $\mathbf{l}^k_{i,j}$ and $\mathbf{f0}^k_{i,j}$. Then, we defined scores based on F0, loudness and duration for a term i in a segment j, by combining the values coming from the multiple occurrences of the term, as follows:

$$\mathrm{f0}(i,j) = \max_k\{\max(\mathbf{f0}^k_{i,j})\} \quad \text{(Pitch)}$$

$$\text{f0-range}(i,j) = \max_k\{\max(\mathbf{f0}^k_{i,j})\} - \min_k\{\min(\mathbf{f0}^k_{i,j})\} \quad \text{(Pitch Range)}$$

$$\mathrm{l}(i,j) = \max_k\{\max(\mathbf{l}^k_{i,j})\} \quad \text{(Loudness)}$$

$$\mathrm{d}(i,j) = \max_k\{\mathrm{d}^k_{i,j}\} \quad \text{(Duration)}$$

Finally, we combined the previous scores Pitch, Loudness, Duration, Pitch Range in order to explore different definitions for the final acoustic score of a term in a segment. In particular, we experimented with the following six definitions for $\mathrm{ac}(i,j)$:

- The individual scores:

$$\mathrm{ac}(i,j) = \{\mathrm{f0}(i,j), \text{f0-range}(i,j), \mathrm{l}(i,j), \mathrm{d}(i,j)\}$$

- The product of Pitch Range and Loudness:

$$\mathrm{ac}(i,j) = \text{f0-range}(i,j) * \mathrm{l}(i,j)$$

- The product of Pitch and Loudness:

$$\mathrm{ac}(i,j) = \mathrm{f0}(i,j) * \mathrm{l}(i,j)$$

3.3 Indexing and Retrieval

The text transcripts of each slide group segment were indexed with the Terrier IR platform[5] [10] version 3.5. We extended Terrier to also store the prosodic features associated with term occurrences in the inverted index along with the standard IR term frequency statistics. In addition, we setup Terrier to properly handle Unicode strings by setting its tokeniser class to *UTFTokeniser*, the property *trec.encoding* to *utf-8* and the property *string.use_utf* to true.

Retrieval was performed with Terrier, with an extension of the weighting model implemented by the class *TF_IDF*. When using this weighting model, Terrier computes relevance scores by following the probabilistic approach, in which the relevance score function is defined as the sum of individual term weights as presented in Equation 2.

Equation 3, along with the following definitions for $\mathrm{tf}(i,j)$ and $\mathrm{idf}(i,C)$, based on the Okapi BM25 weighting function [13], are implemented by the class *TF_IDF* in Terrier to complete Equation 2:

$$\mathrm{tf}(i,j) = \frac{k_1 * \mathrm{tf}_{i,j}}{\mathrm{tf}_{i,j} + k_1 * (1 - b + b * \frac{\mathrm{dl}_j}{\mathrm{avdl}_j})} \quad (9)$$

$$\mathrm{idf}(i,C) = \log(\frac{N}{n_i} + 1) \quad (10)$$

In Equation 9, $\mathrm{tf}_{i,j}$ is the number of occurrences of term i in segment $\mathbf{s}_j$, while dl_j, and avdl_j are the number of terms in $\mathbf{s}_j$ and the average length from all the segments in C respectively. In Equation 10, N is the total number of segments in C while n_i represents the number of segments in C containing the term i.

We extended the *TF_IDF* weighting model implemented in Terrier by integrating terms' acoustic scores to the weighting function. We experimented with the harmonic mean given in Equation 7 and with the following weighted linear interpolation:

$$\mathrm{w}(i,j) = \mathrm{idf}(i,C) * (\alpha * \mathrm{tf}(i,j) + (1-\alpha) * \mathrm{ac}(i,j)) \quad (11)$$

In both weighting functions, $\mathrm{tf}(i,j)$ and $\mathrm{idf}(i,C)$ were implemented as in the original *TF_IDF* model from Terrier (Equations 9 and 10) and $\mathrm{ac}(i,j)$ was always one of the six acoustic scores that we presented in Section 3.2. Overall, we experimented with 12 weighting functions, each of which defined a different "prosodic-based" retrieval model.

Note that the weighted linear interpolation from Equation 11 does not combine acoustic scores with inverse document frequencies. The reason for this comes from the assumption that the acoustic score $\mathrm{ac}(i,j)$, which is intended to capture the grade of prominence of a term relative to others in the lecture, is a measure of the importance that the term i has for the segment $\mathbf{s}_j$. The term frequency $\mathrm{tf}(i,j)$ can be also considered a measure of the level of importance of the term i for the segment $\mathbf{s}_j$. Thus, it sounds logical to only interpolate $\mathrm{ac}(i,j)$ with $\mathrm{tf}(i,j)$ in the calculation of the weight $\mathrm{w}(i,j)$.

3.4 Optimisation and Model Selection

In order to optimise the parameters α, θ_{ir}, and θ_{ac} in the prosodic-based weighting functions, we used as our training set the text queries and relevance assessment data from the passage retrieval subtask at the NTCIR-10 "2nd round of IR for Spoken Documents" (SpokenDoc-2) [1] task. The relevance assessment data of the SpokenDoc-2 passage retrieval task associates each query with a sequence of relevant IPUs, as opposed to the relevance data of this year's slide group segment task, in which each query is associated with a slide group ID. So, we had to map each slide group ID to its IPUs sequence in the retrieval results of our training experiments to be able to evaluate these with the relevance assessment data from the SpokenDoc-2 passage retrieval task.

Parameter optimisation was performed by evaluating the prosodic-based weighting functions with different parameter values. Then we selected the retrieval models that obtained the best results in terms of the evaluation metrics used in the SpokenDoc-2 passage retrieval task. We evaluated the weighted linear interpolation from Equation 11 by varying α from 1.0 to 9.0 and the harmonic mean from Equation 7 by varying both θ_{ir} and θ_{ac} between 1.0 and 5.0. In addition, we evaluated Terrier's retrieval model as implemented by the class *TF_IDF* and considered this as our baseline system.

[5] `http://terrier.org`

In an attempt to avoid overfitting when selecting the best performing models, we split the query set from the SpokenDoc-2 passage retrieval dataset into two sets, $SD2_1$ and $SD2_2$, and evaluated the prosodic-based models on each of these. We then ranked the models according to their performance in terms of Utterance MAP (uMAP) for $SD2_1$ and $SD2_2$, and selected the ones that performed better based on both ranks. To do this, we simply selected the models with the highest sum of rank numbers in $SD2_1$ and $SD2_2$. This procedure was performed individually for each type of transcripts to obtain different best performing models for spoken segments that were transcribed under the manual, *math*, and *unmatchAMLM* conditions. Note that in the SpokenDoc-2 tasks, the query sets only included text queries, so our best models were selected on the basis of their performance with this type of queries.

Table 1 shows results for some of the best performing models obtained from our model selection procedure. The results presented in the table were obtained by evaluating the best performing models with the complete query set from the SpokenDoc-2 passage retrieval task, this included the queries from the two parts $SD2_1$ and $SD2_2$. In the table, we sorted the retrieval models according to their performance in uMAP and not according t othe sum of their rank numbers in $SD2_1$ and $SD2_2$, so Table 1 may not reflect the real ranking order obtained during our model selection process. We also marked in bold the best performing models that were finally submitted for evaluation at the SQ-SCR task. Note that we submitted runs with our baseline model as well, in order to compare its performance against the prosodic-based models. The k_1 and b parameters in Equation 9 were set to the default values of 1.2 and 0.75 respectively.

4. EXPERIMENTAL RESULTS

Figures 1, 2, 3 show the evaluation results for the runs using prosodic-based models and the baseline system on the SQ-SCR slide group segment task. The barplot from Figure 1 shows the results obtained when the manual transcrips of the segments are used, while the barplots in Figures 2 and 3 show the results obtained when the ASR *match* and *unmatch* transcripts of the segments are used.

The barplots show the MAP value obtained for each retrieval model with an individual bar. We use the same patterns across barplots for idenfitying the same or similar models, e.g., models based on LI-LPr have a wide diagonal strip patter in all the barplots. Bars are grouped depending on the type of spoken query used. So, in each barplot, bars from the "Manual" group are the results of models evaluated with the manual transcription of the spoken queries, whereas the "Match" and "UnmatchAMLM" bar groups show results for models evaluated with the ASR *match* and *unmatchAMLM* transcripts of the spoken queries.

A general trend can be noted is that the prosodic-based retrieval models did not perform better than the baseline model TF_IDF. An exception is the case where manual transcripts are used for both spoken queries and segments (lefmost bar group in Figure 1). In this case, the models LI-Pr-0.7 and LI-LPr-0.7 obtained MAP values of 0.121 and 0.110 respectively, which are slightly better than the 0.108 obtained by the TF_IDF baseline. Here, it is important to note that LI-Pr-0.7 and LI-LPr-0.7 were both optimised to perform well over text queries and manual transcripts of segments, as explained in Section 3.4 (see also Table 1). This apparent superiority of LI-Pr-0.7 and LI-LPr-0.7 over the baseline disappears when any other transcript type is used for either the spoken queries or the segments. The remainder of the prosodic-based models shown in the barplots in Figures 2 and 3 were optimised for text queries and the *match* and *unmatchAMLM* type of transcripts, respectively. These barplots show that the models do not beat the baseline under these evaluation conditions. However, there is still an open question of whether these models can outperform the baseline when the manually transcribed spoken queries are used as input instead of their ASR counterparts.

While the raw MAP values may suggest that some retrieval models are more effective than others, paired t student's statistical significance tests considering a 95% confidence level show that, there is no statistical significant difference between the models when they are evaluated over the same combination of spoken queries and segment transcripts. This is, there is no statistical significant difference between any combination of results taken from the same bar group in the barplots.

When comparing the performance of the models by varying the type of spoken query used and leaving the type of transcript fixed, we found statistical significant differences in some cases. For instance, this is the case for the model LI-Pr-0.7 in Figure 1 when it is evaluated with manual queries and ASR queries. Also, the same model performs significally worse when it is evaluated over ASR *match* segments with *unmatchAMLM* queries, than when it is evaluated with manual queries (Figure 2). This shows, as expected, that the quality in the transcription of the spoken queries matters. Similarly, the quality of the transcripts used for the segments affects the retrieval results. When using low quality ASR transcripts the results achieved by the models are consistently lower. For example, the results obtained with LI-LPr-0.5 for the *match* segments are statistically significantly lower than the ones obtained for the *unmatchAMLM* segments (Figures 2, and 3).

We also analysed the performance of our models on individual queries. This was done in order to identify cases for which prosodic-based models may have performed well. The barplot from Figure 4 compares average precision (AveP) values obtained for query 1 for every possible combination of spoken query and segment transcript types. Bars with a diagonal pattern show the average AveP obtained by the prosodic-based models for query 1, while solid grey bars show the AveP obtained by the baseline TF_IDF model for query 1. From the barplot, it can be seen that prosodic-based models performed better than the baseline independently on the quality of the transcripts.

5. CONCLUSIONS AND FURTHER WORK

This paper described DCU's participation at the NTCIR-11 SpokenQuery&Doc task. Following on from previous research, we experimented with various weighting functions that attempt to exploit the prosodic prominence of terms in order to enhance their TF-IDF scores.

We participated in the slide group segment retrieval subtask. For the text transcripts provided by the organisers we computed a set of normalised prosodic features for each recognised word and aligned these to the processed manual and ASR transcripts based on the word's timestamps. Transcripts enriched with the prosodic features were then indexed with the Terrier IR platform, and term weighting

Transcript Type	Model ID	Weighting Function	Acoustic Score	Parameters			Results		
				α	θ_{ir}	θ_{ac}	uMAP	pwMAP	fMAP
Manual	**LI-Pr-0.7**	Eq. 11	Pitch Range	0.7	-	-	0.1369	0.0951	0.0995
	LI-LPr-0.7	Eq. 11	Loudness * Pitch Range	0.7	-	-	0.1369	0.0976	0.1005
	G-LP-1-1	Eq. 7	Loudness * Pitch	-	1	1	0.1326	0.0960	0.0989
	BM25	Eq. 3	-	-	-	-	0.1270	0.0950	0.0972
Match	**LI-LPr-0.5**	Eq. 11	Loudness * Pitch Range	0.5	-	-	0.0842	0.0508	0.0524
	LI-Dur-0.3	Eq. 11	Duration	0.3	-	-	0.0819	0.0498	0.0521
	G-Pr-1-1	Eq. 7	Pitch Range	-	1	1	0.0786	0.0473	0.0499
	LI-Pr-0.7	Eq. 11	Pitch Range	0.7	-	-	0.0778	0.0490	0.0501
	BM25	Eq. 3	-	-	-	-	0.0682	0.0477	0.0486
UnmatchAMLM	G-P-3-1	Eq. 7	Pitch	-	3	1	0.0288	0.0208	0.0131
	LI-LP-0.5	Eq. 11	Loudness * Pitch	0.5	-	-	0.0278	0.0210	0.0135
	LI-LPr-0.2	Eq. 11	Loudness * Pitch Range	0.2	-	-	0.0271	0.0205	0.0132
	LI-P-0.9	Eq. 11	Pitch	0.9	-	-	0.0227	0.0206	0.0129
	BM25	Eq. 3	-	-	-	-	0.0222	0.0203	0.0128

Table 1: Evaluation results on the SpokenDoc-2 dataset for some of the best performing retrieval models. The last columns show the results in terms of Utterance-based MAP [uMAP], Pointwise MAP [pwMAP], and Fractional MAP [fMAP]. Retrieval models marked in bold in column "Model ID" were submitted for evaluation at the SQ-SCR task.

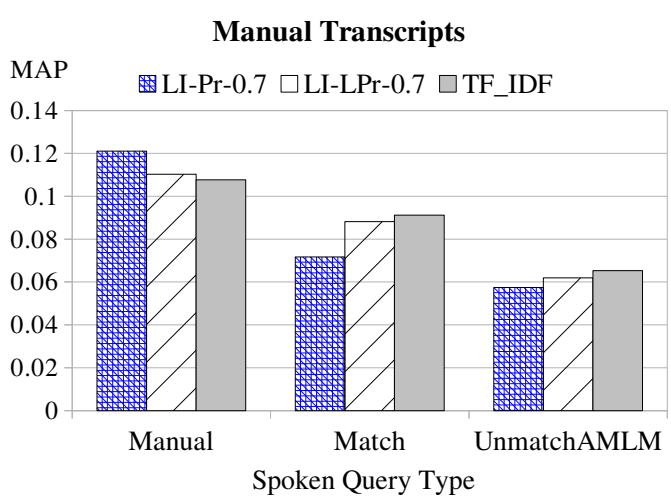

Figure 1: Evaluation results for the case when the manual transcriptions of the segments are used.

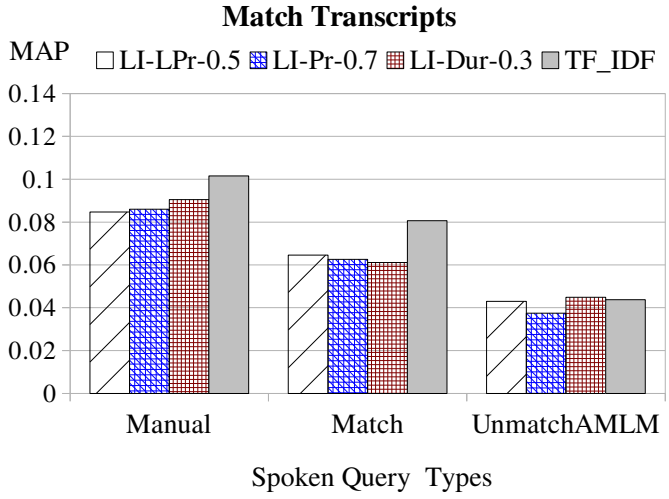

Figure 2: Evaluation results for the case when the ASR *match* transcripts are used.

schemes were implemented to combine the prosodic information with TF-IDF scores. We submitted for evaluation the prosodic-based retrieval models that obtained the highest uMAP when evaluated on the SpokenDoc-2 query set.

The evaluation results do not provide sufficient evidence to conclude that our prosodic-based retrieval models improve over a simple baseline. It thus remains an open question whether prosodic prominence at the word level can be effectively used to improve retrieval performance in the Spoken-Query&Doc task. However, as we would expect, the results show that transcript quality of both spoken queries and segments impacts on the retrieval effectiveness of the models. The results do though suggest that the prosodic-based models may be useful in particular cases, this is supported by the fact that the prosodic-models retrieved more relevant segments at higher ranks than the baseline for some queries. Understanding the situations in which prosodic-models improve retrieval effectiveness, and seeking to generalize these effects will be a subject of our further work.

6. ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL Centre for Global Intelligent Content (CNGL II) project at Dublin City University.

7. REFERENCES

[1] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, and H. Nanjo. Overview of the NTCIR-10 SpokenDoc-2 Task. In *Proceedings of the NTCIR-10 Workshop Meeting*, pages 573–587, Tokyo, Japan, 2013.

[2] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the NTCIR-11 SpokenQuery&Doc task.

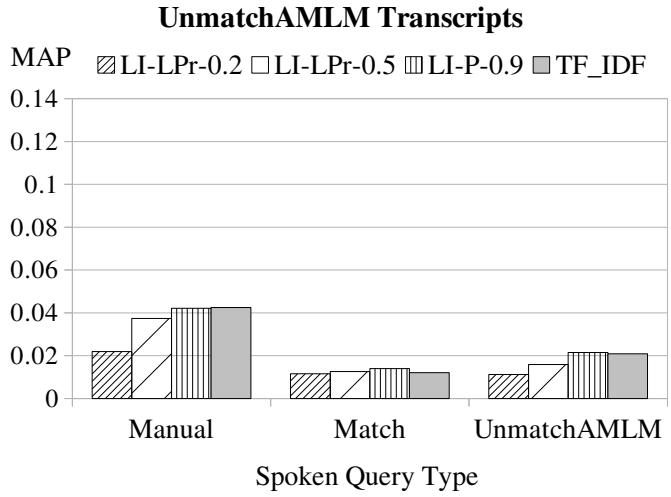

Figure 3: Evaluation results for the case when the ASR *unmatchAMLM* transcripts are used.

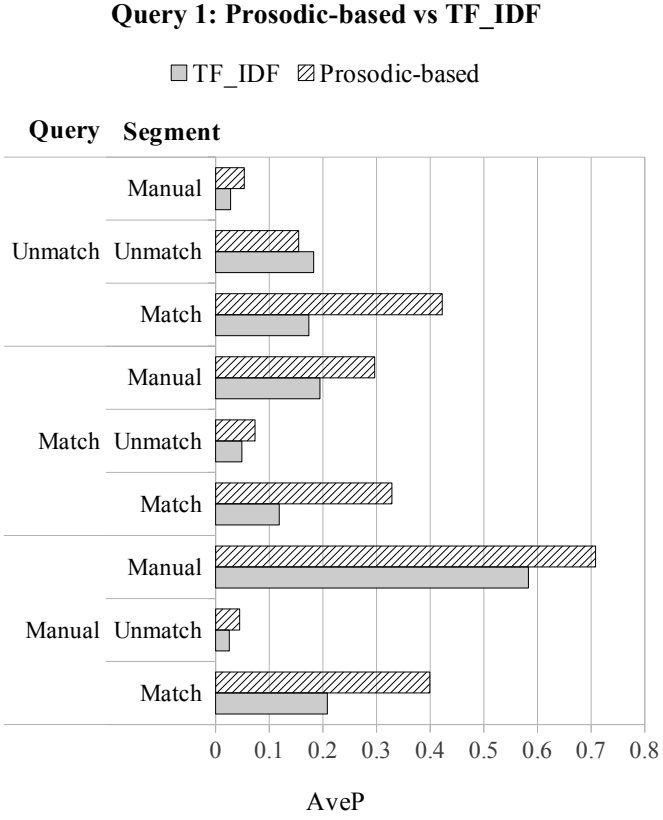

Figure 4: Average Precision [AveP] of prosodic-based models (averaged) versus average precision of TF_IDF for Query 1.

In *Proceedings of the NTCIR-11 Conference*, Tokyo, Japan, 2014.

[3] B. Chen, H.-M. Wang, and L.-S. Lee. Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *Proceedings Interspeech'01*, pages 299–302, Aalborg, Denmark, 2001.

[4] F. Crestani. Towards the use of prosodic information for spoken document retrieval. In *Proceedings ACM SIGIR'01*, pages 420–421, New Orleans, USA, 2001.

[5] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at MediaEval 2014. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[6] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In ACM, editor, *Proceedings ACM Multimedia (MM)*, pages 835–838, Barcelona, Spain, October 2013.

[7] C. Guinaudeau and J. Hirschberg. Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. In *Proceedings Interspeech'11*, pages 1401–1404, Florence, Italy, 2011.

[8] J. Hirschberg. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1):31–43, 2002.

[9] I. Lehiste. *Suprasegmentals*. M.I.T. Press, 1970.

[10] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in Terrier: a search engine for advanced retrieval on the web. *Novatica/UPGRADE Special Issue on Next Generation Web Search*, pages 49–56, 2007.

[11] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings ACM SIGIR'98*, pages 275–281, Melbourne, Australia, 1998. ACM.

[12] D. N. Racca, M. Eskevich, and G. J. F. Jones. DCU search runs at MediaEval 2014 Search and Hyperlinking. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[13] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST Special Publication 500-225, 1995.

[14] A. Rosenberg. AuToBI - a tool for automatic ToBI annotation. In *Proceedings Interspeech'10*, pages 146–149, Makuhari, Japan, 2010. ISCA.

[15] G. Salton. Mathematics and information retrieval. *Journal of Documentation*, 35(1):1–29, 1979.

[16] R. Silipo and F. Crestani. Prosodic stress and topic detection in spoken sentences. In *Proceedings SPIRE'00*, pages 243–252, A Coruña, Spain, 2000.

[17] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808, 2000.

[18] M. Wagner and D. G. Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945, 2010.