Journal of Engineering and Applied Sciences 9 (10-12): 457-463, 2014
ISSN: 1816-949X

An Enhanced Phoneme-Matching Algorithm Enhanced by User Feedback to Identify Possible Automatic Speech Recognition Transcription Errors

James Carmichael
Al Ghurair University, Dubai, UAE

Abstract: This study reports on recent improvements made to a Phoneme-Matching Algorithm (PMA) reported in a previous study. Similar to its predecessor, the purpose of the Enhanced PMA (EPMA) is to identify word recognition errors in automatically generated transcripts detailing the speech content of digital multimedia soundtracks that are routinely queried by professional researchers (such as academics and archivists). In order to alert a user to the possibility that a particular search term may have been incorrectly recognised as some other word or phrase, the EPMA when invoked during a query operation will parse the transcript's text to locate words or phrases of similar phonetic structure to the query term and then present these suspected speech recognition errors to the user for consideration. The EPMA'sperformance has been improved by incorporating techniques to learn from user feedback concerning error identification. When tested on a corpus of digital multimedia, the EPMA averaged an 80.55% success rate in correctly identifying words/phrases which were actually instances of misrecognised query terms.

Key words: Automatic speech recognition, word error rate, speech-to-text transcripts, query terms, identifying

INTRODUCTION

The thematic indexing of European Cultural Heritage (CH) multimedia documents constituted one of the key objectives of MultiMatch (http://www.multimatch.org), an EU-funded research project which produced as its principal deliverable a multilingual, multimodal web-based Information Retrieval (IR) search engine. This search engine represents a successful proof-of-concept implementation demonstrating that access to digital information need not be hindered by barriers of language or document type; it should not be the case for example, that a search operation generated by a Dutch-language query is unlikely to consider Spanish-language documents (Marlow *et al*., 2008), particularly if their associated metadata descriptors are also written in Spanish. To counteract such a bias against Cross Language IR (CLIR), the MultiMatch search engine provides bi-directional translation services supporting six European languages (i.e., English, Dutch, Italian, Spanish, Polish and German); the intention here is that a query written in any of these languages can be automatically translated into all of the other five and the resulting query translations are then used to search the entire MultiMatch digital library of both multimedia and text (html) documents. Furthermore, it is possible to search this digital library using a variety of multimodal query techniques developed specifically for multimedia intra-document searching (i.e., searching for a specific segment of interest within a document). However, the technology underpinning such multimedia-based document searching is still somewhat fallible: in a previous study investigating the effect of speech transcription inaccuracies on IR operations, Carmichael (2008) found that incorrect Automatic Speech Recognition (ASR) transcriptions of key search terms resulted in a more than four-fold increase in the average time taken by users to locate specific spoken instances of search terms on selected video soundtracks. The following section describes certain improvements made to the design and implementation of a software application which identifies possible ASR transcription errors based on the assumption that the user-supplied search term may have been misrecognised as another phonetically similar word or phrase.

MATERIALS AND METHODS

Enhancing the "nearest-neighbour" phoneme matching algorithm: As discussed in the previous study, incorrect ASR processing of a multimedia document's soundtrack has proven to be a substantial hindrance for users when searching multimedia digital libraries for sub-document segments (i.e., audio or video clips). The Phoneme Matching Algorithim (PMA) developed by Carmichael (2011) represents an early attempt to counteract the effect of such speech-to-text transcription errors; this study reports on an attempt to enhance the PMA's

error-spotting capability by incorporating a feedback mechanism which allows the user to manually identify speech recognition errors and then suggest corrections to said errors. The implementation of the Enhanced PMA (EPMA) is described in the subsection that follows.

Architecture of the EPMA: One of the core operations of ASR speech processing is the mapping of speech sounds to specific phonemes for a given language. It is usually the case that instances of incorrect sound-to-phoneme mapping are to some degree affected by speaking style and prosody (Rabiner and Juang, 1993). Nevertheless, a previous study by Carmichael (2011) notes that there are certain types of commonly-occurring misrecognition errors which are not specific to any particular speaker or group of speakers. It is noted for example that voiced and unvoiced consonant pairs (e.g., /d/ and /t/, /p/ and /b/) are more likely to be confused one with another as opposed to being confused with other speech sounds that are produced using substantially different articulatory movements.

It is also not uncommon for ASR parsing to incorrectly delineate word boundaries in fluent connected speech an error which often results in word insertions and/or deletions (e.g., erroneously transcribing "looking" as "look in"). The aforementioned types of errors share one common feature: the ASR techniques applied have successfully captured the most salient phonetic features of the utterance but were less accurate at achieving a more fine-grained level of semantic analysis. The phoneme matching algorithm works on the assumption that the majority of misrecognition errors produce phoneme confusions which share the same general phonetic characteristics. It is therefore, possible to systematically detect these errors solely via the application of text-based Natural Language Processing (NLP) techniques.

As can be expected, the adoption of a primarily text-based approach to correcting speech processing errors presents its own set of challenges: since the analytical techniques employed do not reference the original acoustic data which generated the speech transcriptions, the phonetic composition of any selected word or phrase is determined in the first instance by reference to its orthography. Of course, it is not always possible to deduce a word's phonemic structure based solely on the manner in which it is spelt. The spelling conventions and pronunciation rules of many conventional alphabets tend to exhibit some degree of ambiguity and inconsistency in relation to the mapping between the parent language's speech sounds and various characters in its alphabet. In the case of the English language for example, the letter combinations "u", "oo" and "ou" could all be used to represent the phone/u:/ while "u" and "oo" can also designate the schwa/ə/ (thus the "oo" letter sequence in "look" is realised as/ə/ while the same letter combination in "loot" is rendered as /u:/). When parsing speech transcripts, the EPMA retrieves information about the phonetic structure of any given word by accessing a look-up table containing a list of approximately 10,000 words and their corresponding phonetic composition, including any major pronunciation variation thereof for example, the word "to" has two phonetic transcription possibilities, either /tu:/ or /tə/. In the event that an unknown word (i.e., an Out Of Vocabulary [OOV] item) is encountered in the ASR transcription being processed, the system will prompt the user to provide a pronunciation of the word via the use of a speech synthesiser which concatenates manually selected phoneme sequences in order to produce a word-level text-to-speech pronunciation which the user then verifies or rejects. If necessary, alternate pronunciations may also be specified for the word/word sequence of interest.

After prompting for and being supplied with the phonetic details of any OOV items, the EPMA parses the selected speech transcript, converting all words therein to their phonetic transcriptions, the resulting phoneme sequences are then compared with those of the user-supplied query term to determine if there are any sequences in the target speech transcription which due to phonetic similarity are likely to be misrecognised instances of the query term. Phonetic similarity may be evaluated using place and manner of articulation as a criterion: when comparing a phoneme from the query term with one from some word in the target speech transcription, phonetic similarity is determined via an analysis of the place and manner of articulation (Ladefoged, 1982) of the respective phones. For example, /a/ is more likely to be confused with /æ/ than with /i/, since, /i/ is high frontal vowel (i.e., its pronunciation requires the tongue to be in an elevated position towards the front of the mouth) while /a/ and /æ/ are low back vowels. Accordingly, the EPMA incorporates a classification scheme which assigns numerical values to phonemes according to their place and manner of articulation, so that the lowest scores are assigned to speech sounds produced towards the back of the oral cavity with the tongue depressed; conversely, the phones produced with the tongue raised and towards the front of the mouth are accorded higher values. In addition to the foregoing distinctions, vowels due to their longer duration are assigned a greater weighting than consonants (Table 1 and 2).

Table 1: EPMA scoring table for English-Language Consonants

Consonants	Values
g (get)	1
n (Sing)	2
k (Kate)	3
ɹ (radio)	4
h (hotel)	5
j (yes)	6
ʤ (jaw)	7
tʃ (chew)	8
ʒ (Treasure)	9
ʃ (ship)	10
j (yes)	11
l (late)	12
z (zebra)	13
s (silly)	14
d (date)	15
t (tea)	16
ə (these)	17
ɵ (thin)	18
f (fox)	19
b (big)	20
p (pig)	21
n (night)	22
m (man)	23
w (wet)	24

Table 2: EPMA scoring table for english language vowels, diphthongs and triphthongs

Vowe	Values
ɒ (Pot)	25
ɔ: (thaw)	26
ɔɪ (void)	27
ɑ: (father)	28
æ (cat)	29
ɑɪ (dive)	30
ɑɪə (fire)	31
ɑʊ (out)	32
ɑʊə (flour)	33
oʊ (hoe)	34
ʊə (poor)	35
ə (alone)	36
Λ (cut)	37
ʊ (pull)	38
ɜ: (burn)	39
eɪ (paid)	40
I (build)	41
Iə (beer)	42
ɛ (bet)	43
ɛə (bear)	44
u (zoo)	45
I: (see)	46

This points-based scoring system makes it possible to calculate a Phonetic Divergence Score (PDS) when comparing any two given phoneme series with the difference in the selected phonemes' scores indicating the degree of phonetic divergence, thus the lower the score, the greater the similarity (e.g., a score of '0' would indicate that the two phoneme strings compared are identical).

For the experiments detailed in this study, PDS scores have been standardised by way of dividing the total PDS score by the number of phonemes in the query term, thus a 10-phoneme query term garnering a gross PDS score of 30 would be assigned a standardised PDS score of 3. For the purposes of this investigation, a standardised PDS score of 3.0 or lower indicates that the phoneme string earning such a score is likely to be a misrecognition of the query term. Accordingly, a PDS score of 3.0 has been set as the threshold misrecognition value, i.e., a PDS score equal to or lower than 3.0 indicates that the word or phrase receiving that score will be considered as a query term misrecognition.

The EPMA algorithm also incorporates procedures to compensate for phenomena known as deletion and insertion. Since, a misrecognition of a speech utterance will probably contain phonetic errors, the EPMA's parsing procedure allows some degree of elasticity in order to offset the effects of deletion or insertion due to "nearest neighbour" substitutions (The "nearest neighbour" error is a phenomenon wherein the phoneme sequence extracted from the acoustic signal does not correspond to any of the items in the ASR application's vocabulary, thus the "nearest neighbour" (i.e., word sequence with the most closely matching phoneme sequence) is output instead). To counteract this problem of spurious phonemes, the EPMA employs a 'look-back-then-look-forward' technique when iteratively comparing a query term's phoneme sequence with any given sequence of equal length taken from the target speech transcription (Fig. 1). In this fashion, when the EPMA compares phoneme sequences from the query term and speech transcript respectively, the nth phoneme in the query term's sequence is compared with the n-1st through to the n+1st phoneme of the corresponding ASR transcript to determine which of the three in the series is the most similar phonetically to the phoneme of interest in the query term; if it is observed that the n-1st or the n+1st unit of the corresponding speech transcript phoneme series is indeed the best candidate for the query term's nth phoneme, then said nth phoneme's immediate neighbours are also compared with their speech transcript counterparts to determine if there is a consistent 'out-by-one' mismatch, indicating the presence of an extraneously inserted phoneme in the event that there is good evidence indicating the presence of such phonetic 'noise', then the extraneous phoneme is excluded from further analysis (Table 3).

In addition to the automated text-parsing techniques described above, the EPMA can also improve its performance based on user-supplied corrections of misrecognition errors. In such a scenario, the user is prompted to manually identify and correct instances of misrecognised items. The EPMA will then attempt to establish if these user-spotted errors are the result of

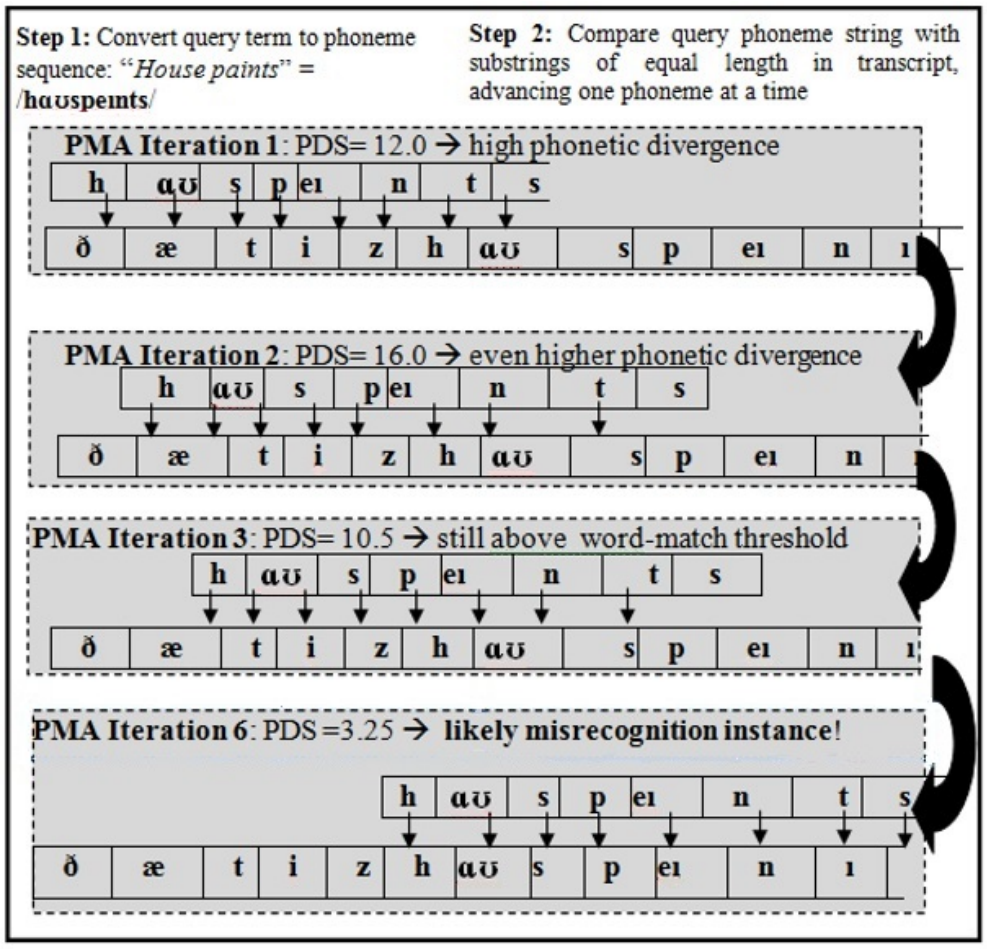

Fig. 1: Example of EPMA phoneme-parsing to find misrecognised query term "house paints"

Table 3: Example of ASR misrecognition errors for a MultiMatch audio podcast

Variables	Discription	Values	Characterstics	Parameters
Correct version	Artists also soon began to use other types of paints developed for industrial use	Such as house	Paints and	Car paints
ASR Transcript (Verbatim)	Artists also soon began to use other types of paints developed for industrial use	That is how	Spain is an	Car paints

Italicised text = an ASR misrecognition of some specific word. Underlined and italicised text = ASR word insertion

some consistent phoneme confusion which is specific to that particular speaker and/or peculiar conditions within the acoustic environment (such as loud background noises).

If it is apparent that certain speakers or acoustic environments are likely to produce specific phonemic confusions, then the details of these phonemic confusions along with the speaker's identity are recorded for the purposes of training/enhancing the EPMA's error-spotting algorithms so as to produce an improved performance if other text-to-speech 11 transcripts from that speaker are encountered on future occasions.

EPMA testing protocols: The audio content of the multimedia documents selected to evaluate the EPMA represent a range of recording conditions and speech content typical of the larger MultiMatch collection. Although, all of the documents in the test corpus feature primarily English-language speech content (to simplify ASR processing), they are still representative of the quality and diversity of speech material for podcasts in other MultiMatch-supported languages, since this test collection contains several documents which exhibit:

- A higher than normal frequency of foreign language items
- Multiple speakers, occasionally speaking simultaneously
- Speech uttered in the presence of certain types of background noise (e.g., music)

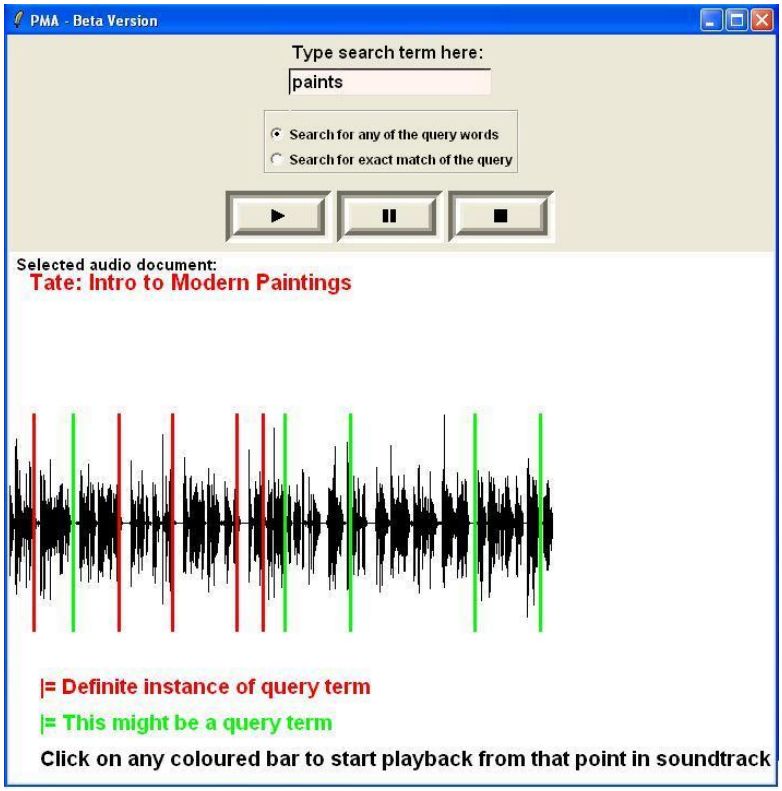

Fig. 2: EPMA user interface (after execution of a query operation)

Each multimedia document in the test corpus which comprised of 50 documents formatted as 11 Khz-sapmled audio files encoded in the wav format was processed by the ASR application to produce a transcript with word-level timestamps. These ASR-generated time-stamped transcripts were then compared with manually generated error-free transcripts (also featuring word-level timestamps) in order to determine the number of speech recognition processing errors and the likelihood that they were misrecognitions of some specific word/phrase or cases of the ASR application mistaking a non-speech sound for meaningful speech (Fig. 2).

For the purpose of making the testing exercise manageable and approximating as closely as possible real-life search scenarios, a series of pre-defined queries as formulated by a group of CH professionals were suggested for each multimedia document of the test group. In order to accomplish the two-fold objective of:

- Standardising the queries in terms of syllable length
- Investigate the effect of a query's phonetic complexity on the EPMA's performance

The pre-defined 100 query items were categorised as either short queries (up to four syllables) or long queries (5-8 syllables) with none of the queries exceeding eight syllables. In the case of test corpus document ID#034 for example, the three suggested queries were as follows: "Painting" (2 syllables), "Synthetic paints" (4 syllables), "pre-nineteenth century" (6 syllables). It must be noted that the aforementioned eight-syllable limit does not imply that multi-word queries containing phrases/sentences totalling >8 syllables were not permitted, however the EPMA is configured to process such multi-word query items as if they were a series of single-word queries, e.g., the query "Impressionist movement" would be executed as two separate queries unless the user specified that the query phrase should be treated as an indivisible unit due to some specific connotation attached thereto, e.g., "Golden Age" or "Fifth Symphony". The EPMA's performance was measured according to two criteria namel.

The Percentage of Correct Identifications (PCI) of misrecognised words/phrases which are actually instances of some user-generated query term; this percentage is calculated based on the total number of search term misrecognition instances (as determined by a

manual inspection) present in the selected transcript. For example, a correct identification score of 80% would indicate that the PMA correctly identified eight out of a total of ten query term misrecognitions for a given speech transcript.

The percentage of false positives generated by the EPMA (a false positive is defined as the declaration of a word or phrase in the transcript to be a misrecognition instance of a query term when such is not the case). This false positive percentage is calculated based on the total number of words appearing in the selected transcript. For example if the PMA erroneously declared 100 words in a 200-word transcript to be instances of a query term, then the false positive percentage would be 50%. It must also be noted that the testing of the PMA was largely an automatic process and only involved the direct participation of a user-group for the purpose of manually correcting the speech-to-text transcripts.

RESULTS AND DISCUSSION

EPMA test performance results: A total of 100 queries half of which were four syllables or less while the rest were between five and were conducted for the fifty selected multimedia documents. Figure 3 depicts the EPMA's identification accuracy when processing these queries. In terms of correct identification of misrecognised query terms, the PMA's performance ranged from 44.6-100% with a mean average of 73.64% and a modal (i.e., most frequent) average of 72.2%. As expected, the EPMA's accuracy is better for queries exceeding 4 syllables. In terms of performance particulars, it was observed that as expected the EPMA's capacity to detect misrecognition errors was degraded whenever the performance of the ASR engine itself deteriorated as a result of processing non-English speech and/or speech featuring multiple talkers active simultaneously. It would appear that such conditions increased the incidence of phonetically random misrecognition errors, i.e., errors with PDS values exceeding 5, indicating only a very limited phonetic resemblance to the original utterance. As a result, the EPMA on the six occasions when it encountered speech transcript segments representing such multi-speaker conditions only averaged 21.4% for correct identification of misrecognition instances. Regrettably, the improvements in the EPMA's architecture still did not enable it to detect any misrecognition instances in conditions where speech occurred in the presence of intrusive background noise.

In terms of the false positive evaluation criterion, the PMA's overall average for the test corpus was 0.6% for 5-8 syllable queries which corresponds to the PMA

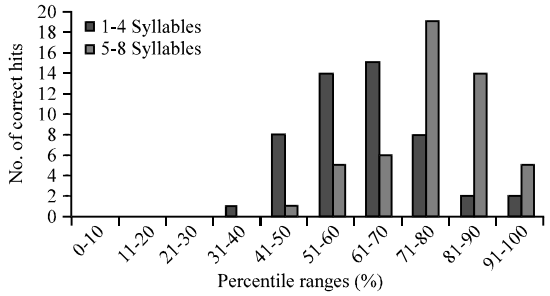

Fig. 3: Comparison of EPMA performance (measured in correct identification of ASR misrecognitions of query term) for 1-4 syllable and 5-8 syllable queries

erroneously flagging as query term misrecognitions between one word per 100 in the selected speech transcripts; for short queries (1-4 syllables), the incidence of false positive was 1.4%, more than twice that of the long queries, a result which is consistent with expectations since, long queries tend to contain more phonemes an increase in phonetic complexity which lessens the likelihood of a false positive.

Provision of manually corrected data to improve EPMA performance: In order to improve a given ASR application's performance for a specific individual, the usual procedure is to provide said ASR application with a range of speech samples from the target speaker so that the application can analyse the "training" speech data and thereforebetter adapt to that person's speaking style (Rabiner and Juang, 1993). Unfortunately, constraints of time and resources did not permit the application of any speaker adaptation methods during the ASR processing of the MultiMatchspeech data used for this experiment. Instead, an indirect speaker adaptation method was attempted whereby the EPMA for all instances where it failed to detect ASR misrecognition errors was provided with information which included:

- The manually corrected version of the misrecognised item (For example if the ASR application misrecognised car" as "for" in a specific speech transcript, then the EPMA was provided with both the correct and misrecognised transcriptions of the speaker's utterance along with the actual identity of said speaker)
- The identity of the speaker who uttered the misrecognised word/phrase
- The specific type of background noise present when the misrecognised item was spoken

In order to ascertain if the EPMA's performance benefitted from this novel approach to speaker adaptation

training, the manual corrections made to three speech-to-text transcripts produced by the same speaker were processed by the EPMA for training purposes and then two other scripts (produced by the same speaker but not present in the training data set) were used for testing purposes. This training-testing procedure was repeated for transcripts provided bythree speakers whose speech data also featured in the larger corpus used in the first experiment described in this study.

This researcher is pleased to report that the speaker adaptation procedure detailed above resulted in a 21.50% improvement in the EPMA's error-spotting capacity for the selected individuals' speech data in particular and an 8.35% improvement overall for the larger speech data corpus (This 8.35% improvement translates to an overall EPMA performance upgrade from 72.2% (without the benefit of speaker adaptation) to 80.55%).

CONCLUSION

The EPMA's improved success rate (when compared to the earlier version of the PMA reported by Carmichael (2011) in identifying misrecognised query terms suggests that the look-back-look-forward technique to compensate for phone insertion is a valid ASR error detection method. Such encouraging results notwithstanding, a more thorough confirmation of the utility of this approach would require more extensive and varied testing, including the involvement of human participants in future evaluation procedures, so as to assess how their behaviour is modified during multimodal information retrieval tasks. The introduction of such user-centred evaluation is of pivotal importance since it directly addresses the principal purpose of the PMA which is to sensitise the user to the fallibility of automatic speech recognition systems and by so doing, inspire a more judicious use of all available search modalities in any given IR scenario. It would defeat the purpose for example if the user were to assume that the PMA was itself infallible, i.e., that the PMA will consistently detect under any conditions all ASR misrecognitions of a given query term. To this end, it may be useful to establish the relationship between the PMA's capacity to identify word errors and the overall quality of ASR processing for a given speech transcript. In this way, it might be feasible via some form of confidence measure (such as that devised by Cox) based on a speech transcription's word error rate to advise the user of what to expect in terms of the degradation of the PMA's capacity for error detection due to the quantity and quality of noisy text data in the speech transcript.

REFERENCES

Carmichael, J., 2008. Multimedia retrieval in MultiMatch: The impact of speech transcription errors in search behaviour. Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage. Aarhus, Denmark.

Carmichael, J., 2011. An Improved Text-Based Algorithm for Detecting Automatic Speech Recognition Errors: Observations from the Multimatch Project. Arya Bhatta Journal of Mathematics and Informatics, Vol. 3, No. 1.

Marlow, J., P. Clough, N. Ireson, J. Cigarran Recuero, J. Artiles and F. Debole, 2008. The MultiMatch Project: Multilingual/Multimedia Access to Cultural Heritage on the Web, Museums on the Web Conference (MW2008): Proceedings, J. Trant and D. Bearman (Eds.). Toronto: Archives and Museum Informatics.

Ladefoged, P., 1982. A course in Phonetics (2nd Edn.). New York: Harcourt, Brace and Jovanovich.

Rabiner, L. and B.H. Juang, 1993. Fundamentals of Speech Recognition, Prentice Hall PTR, New Jersey.