CONTENT-BASED RECOMMENDER SYSTEMS FOR SPOKEN DOCUMENTS

Jonathan Wintrode[1], *Gregory Sell*[2], *Aren Jansen*[1,2],
Michelle Fox[3], *Daniel Garcia-Romero*[2], *Alan McCree*[2]

[1]Center for Language and Speech Processing, [2]Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD USA
[3]US Department of Defense, Washington DC USA

ABSTRACT

Content-based recommender systems use preference ratings and features that characterize media to model users' interests or information needs for making future recommendations. While previously developed in the music and text domains, we present an initial exploration of content-based recommendation for spoken documents using a corpus of public domain internet audio. Unlike familiar speech technologies of topic identification and spoken document retrieval, our recommendation task requires a more comprehensive notion of document relevance than bags-of-words would supply. Inspired by music recommender systems, we automatically extract a wide variety of content-based features to characterize non-linguistic aspects of the audio such as speaker, language, gender, and environment. To combine these heterogeneous information sources into a single relevance judgement, we evaluate feature, score, and hybrid fusion techniques. Our study provides an essential first exploration of the task and clearly demonstrates the value of a multisource approach over a bag-of-words baseline.

Index Terms— Content-based recommendation, speech retrieval, low resource, i-vectors

1. INTRODUCTION

Within the growing sea of digital media available on the web, automatically connecting consumers with relevant content is an increasingly important goal. Recommender systems attempt to present their *users* with *items* (e.g. merchandise, videos, music, books) of potential interest (see [1] for a good review). When multiple relevance judgments/ratings from a variety of users are not provided for each item, collaborative filtering [2, 3] techniques behind well-known internet recommender systems (e.g. Netflix [4], YouTube [5], and Amazon [6]) do not apply. Instead one must resort to *content-based recommendation* [7, 8] (a.k.a content-based filtering), which attempts to build a personalized model for each user that is independent from other users and relies solely on either manually or automatically derived features characterizing the contents of each item.

Content-based recommendation can be viewed as an information retrieval task where the notion of relevance is more diffuse than finding items pertaining to a simple text query. Instead, the recommendation system must rank novel content using a model of each user's interests or information needs, which can be topically diverse, can be influenced by non-linguistic attributes, and can change over time. In most content-based recommender systems, high-level attributes and metadata (e.g. genre, author, cast, musician) maintained in a knowledge base are featurized to support a more abstract notion of document relations than may be immediately apparent from a simple bag-of-words vector space model [1, 9]. However, these additional features are typically provided manually by database curators [1]. One notable exception are music recommender systems, where a variety of music processing strategies have been employed to *automatically* extract features that characterize properties such as genre, timbre, and tempo [10, 11, 12].

In the speech domain, the closest bodies of related work concern the tasks of spoken document retrieval [13] and topic identification [14, 15]. Both cases most typically rely on speech recognition to tokenize the audio and construct bags-of-words representations for input to subsequent document similarity measures or topic classifiers. In our case, topical analysis is certainly relevant, though a user's interests may be as simple as a single well-defined topic category (sports or politics) or a complex combination of diverse categories. Moreover, speech audio contains a wealth of information beyond the linguistic message that may influence the relevance to a particular user, either explicitly or implicitly. This includes (but is not limited to) the language spoken, the identity of the speaker, and properties of the environment. Thus, in the spirit of the above-mentioned music recommendation systems, we adopt a fully automatic approach that uses (i) low and high resource automatic speech recognition (ASR) to tokenize the message; (ii) acoustic i-vectors [16] to characterize speaker, language and gender; (iii) speech and music activity detection to characterize proportions of each [17]; (iv) zero resource discovery tools to characterize background acoustic events [18]; and, (v) bags-of-subword units to characterize (to varying degrees) all of the above. Given our heterogeneous featurization, we require a back-end combination strategy to produce a single relevance judgment. Toward this end, we investigate early stage (feature-level) fusion, score-level fusion, and a hybrid approach.

Given the novelty of this task to the speech processing community, the standard topic identification and spoken document retrieval evaluations from the past (e.g. [13, 19]) are not sufficient to evaluate the efficacy of our multisource system design. Instead, we assemble a corpus of audio clips extracted from CreativeCommons.org internet videos. Each clip is sourced from one of 22 distinct website collections and we use the collection sources as a proxy for some hypothetical recommender system user's interests (assuming the remaining 21 collections are uninteresting). Critically, each clip contains not only a substantial portion of speech, but also spans a wide variety of acoustic conditions/environments, speaker identities, languages, and topics. However, each collection is largely distinguishable from the others by some combination of the content-based criteria described above. We find that in this evaluation scenario, all information sources provide complementary signal and, when fused, nearly halve the error rate of a bag-of-words baseline. We begin with a complete description of our proposed system architecture.

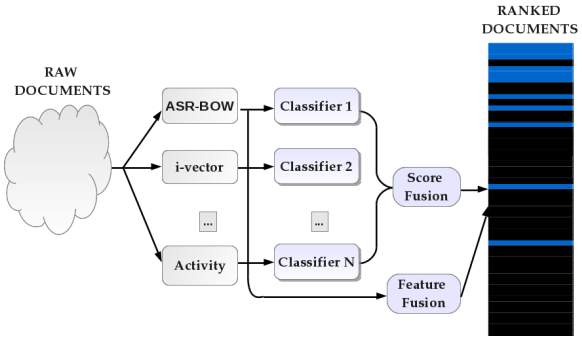

Fig. 1. System architecture

2. SYSTEM ARCHITECTURE

Our recommender system follows the data flow shown in Figure 1. Each document is processed independently, beginning with extraction of each of the feature types described below. Individual classifiers for each feature type are trained, each with the ability to provide its own relevance score and ranking. When considering multiple features simultaneously, we also train back-end fusion classifiers, either fusing scores, the features themselves or a hybrid combination of both. The output of the fusion step also produces a rank order on the test corpus. The training partition is used to train individual feature classifiers, done using a *one-vs-all* approach. For individual features (i-vectors, bags-of-words, etc.) we train both SVM and logistic regression models for each user. We evaluate both classifiers against the development set and use the better model for subsequent fusion experiments.

2.1. Content-Based Features

Each type of feature extraction transforms the audio portion of each document in a corpus to a vector representation. The features range from simple speech duration measurements to bags-of-words from large vocabulary ASR.

2.1.1. English Bags of Words and Senones

We decode each audio file using models built with the Kaldi speech recognition toolkit [20]. Using LDC's Switchboard Phase 1 audio and transcripts [21] we train two models for recognition. The first leverages all 300+ hours of transcribed speech, and the second comprises a randomly selected 10 hour subset. To obtain bags-of-words feature vectors we extract expected unigram counts from the ASR lattices and compute length-normalized TF-IDF (term frequency, inverse document frequency) values for each document. In addition to characterizing word frequencies, past work in speaker, language, and topic identification has demonstrated the value of subword unit frequencies. In this spirit, we use Kaldi deep neural network (DNN) acoustic models to generate expected senone (context-dependent phone HMM states [22]) counts for each document and constructed feature vectors by taking the logarithm of normalized counts as in [22].

2.1.2. Acoustic i-vectors

Acoustic i-vectors were extracted from each audio clip using a 2048-component UBM and 600 dimensional total variability (T) matrix and length-normalized. Hyperparameters were trained on NIST SRE data ('04, '05, '06, '08). Each i-vector was Garcia-Romero normalized before input to the various backends [23].

2.1.3. Speech and Music Activity

Duration of speech [24] and music activity [17] were used as features as well, both in terms of log-compressed duration as well as proportion of overall duration.

2.1.4. Pseudoterms

In addition to bags-of-words, we also extracted bags of *pseudoterms* from each clip using the zero-resource approach detailed in [18]. A pseudoterm can be thought of as a cluster of acoustically similar audio segments and we can compute the same length-normalized TF-IDF values by counting clusters in each audio clip. In particular, they can be repeated words and phrases that function to characterize the topic. However, in a heterogeneous corpus such as this, they also capture repeated spectro-temporal patterns that help characterize the environment. For example, in the video game collections, game sound effects become pseudoterms.

2.1.5. Semantic Features

For both bags-of-words and bags-of-pseudoterms we perform Latent Semantic Analysis (LSA) [25]) to obtain lower-dimensional semantic (or pseudo-semantic, in the latter case) representations of the data. LSA amounts to applying PCA to the bags-of-words vectors. In our case, we consider 10, 100, and 500-dimensional projections computed using randomized PCA [26] on the training vectors and apply the same projection to the development and test sets.

2.2. Fusion

We examined three different fusion regimes for merging the aforementioned array of acoustic and lexical features for recommendation: *score fusion*, *feature fusion*, and *hybrid* combinations of both scores and features. Given N single systems and scores on the development and test data, score fusion is performed by constructing an N-dimensional feature vector from the scores of each system. We train a logistic regression classifier using the development data to predict the relevance judgment, then apply this fusion classifier to the N test scores for each audio clip. For feature fusion we concatenate the feature vectors of individual systems. Unlike score fusion, feature fusion does not require the development corpus and can be trained on the larger training corpus. Finally, our hybrid score and feature fusion also proceeds via concatenation. Fusing N scores with M feature types of dimension $\{d_1, \ldots, d_M\}$ results in a feature vector in the dimension of $N + \sum_i d_i$. Because the classifier score is a component of the hybrid fusion, we restrict ourselves to the development set features plus N model development set scores to train a logistic regression fusion model.

3. DATA

Data consisted entirely of audio stripped from Creative Commons internet videos (see www.creativecommons.org). Candidate videos were secondarily screened to avoid any potentially copyrighted material, after which we identified 22 unique collections of videos to serve as proxies for recommender users. As is the case in the true recommendation scenario, the defining characteristic of each of

Table 1. Content type, number of speakers, and dominant language for each of the 22 collections in the assembled corpus.

#	Content Type	Speakers	Language
1	video games	multi	English
2	video games	single	English
3	video games	single	English
4	news	multi	Romanian
5	programming tutorials	multi	English
6	politics	multi	English
7	public event recordings	multi	English
8	math tutorials	single	English
9	vlog	single	English
10	religion	multi	English
11	interviews	multi	Spanish
12	interviews	multi	English
13	vlog	single	English
14	public service	multi	English
15	sports	multi	English
16	chemistry tutorials	multi	English
17	interviews	multi	English
18	technology	multi	English
19	English lessons	multi	English
20	tourism	multi	English
21	public events	multi	English
22	public events	multi	Chinese

these collections was not known. In some cases, the primary speaker is the dominant connection, while, in others, the topic is the main consistency. Several collections were neither topically nor were restricted to a single speaker or acoustic condition. Table 1 provides primary topic, number of speakers, and dominant language for each of the 22 collections. In addition to the 22 collections, we collected a set of Creative Commons videos that were not connected to any particular collection. These served as as unseen negative examples for all collections. All audio clips were cut off at 5 minutes in duration, summed to a single channel, and downsampled to 8kHz.

Audio clips were divided into train, dev, and test lists according to their creation date. This fits well with the concept of recommendation, in that future decisions must be made based on past labels. The train set for each collection included the 40 oldest clips from that collection as well as the 40 oldest from 11 other collections, plus 696 out-of-set background clips (1176 total, 3.4% positive). The dev set was similarly composed, with the next 20 clips after the training clips from the positive collection along with 20 from the same 11 other collections, plus 265 out-of-set background clips (505 total, 4.0% positive). The test set included the 40 most recent clips from all 22 collections, as well as 1039 out-of-set background clips (1919 total, 2.1% positive).

4. EXPERIMENTS

Our evaluation considered performance of individual features, score-level fusion, feature-level fusion, and hybrid score-feature fusion. Individual one-vs-all classifiers (one per collection per feature) are scored against both development and test sets, used for training and evaluating score fusion. All results are reported on the test partition. To evaluate, we score the test set with each classifier (individual or fusion) and compute the equal error rate (EER) for each topic. Given a sorted list of scores, the EER point occurs at the threshold where P(FA)=P(miss). Unless noted otherwise, we report the average EER across all 22 collections.

Table 2. Average EER (%) for individual features, using both a support vector machine and logistic regression backend.

Feature	SVM	Logistic
Bow300h	13.7	22.3
Bow10h	16.0	25.7
LSA[bow300h]	15.9	15.3
LSA[bow10h]	18.3	18.6
i-vector	10.5	9.8
Pseudoterms	21.8	27.4
LSA[pterms]	18.1	30.0
Speech/music activity	35.1	29.1
Senones	8.6	9.5

4.1. Individual Features

Full results for single features and both classifiers are listed in Table 2. The acoustic i-vectors, a good general-purpose feature for non-topical aspects of the signal, outperform the 300-hour bag-of-words (bow300h) baseline by nearly 4% absolute (13.7% to 9.8%). The effect of 10h ASR system tokenization (bow10h) results in an increase of only 2.3% EER over the baseline. This corresponds to a increase in word error rate (WER) from 33.1% to 46.2% and the proportionally small drop in EER is consistent with previous topic identification results. This is encouraging for low-resource recommender system applications.

Reducing dimensionality with LSA had different effects on bags-of-words versus bags-of-pseudoterms. While the 10 and 100-dimension projections significantly underperform their respective baselines, the relative impact of reducing the feature space to 500 principal dimensions from 30,246 (words) or 1,340,841 (pseudoterm acoustic events) resulted in only a 2% increase in EER for the bags-of-words and a 3.7% *decrease* for the pseudoterms. Given this result, all references to LSA in the performance tables/figures correspond to the 500-dimension projection.

Moving from a low-resource (10 hour ASR) to zero-resource secnario, we note the difference in EER between the bag-of-words features and LSA-reduced zero-resource pseudoterms is only 2.1%, a 4.4% overall degredation from the 300 hour baseline. The speech and music activity performed poorly in isolation (29% EER) but we will see it nonetheless provides useful information for fusion.

Finally, the bags-of-senones produce the best individual feature performance of 8.6% EER.

4.2. Fusion

For score fusion, we selected either the SVM or logistic regression scores given their performance on the development data. Logistic regression is used to combine the individual scores in all cases. For feature fusion, SVM classifiers broadly outperformed logistic regression and are used in all cases. Figure 2 compares score and feature fusion performance for various system combination. Overall, score fusion proved most effective, reducing average EER from 13.7% to 7.8% (5.9% absolute, 43% relative). Interestingly, bags-of-senones provide no additional gain to score fusion, despite yielding the best

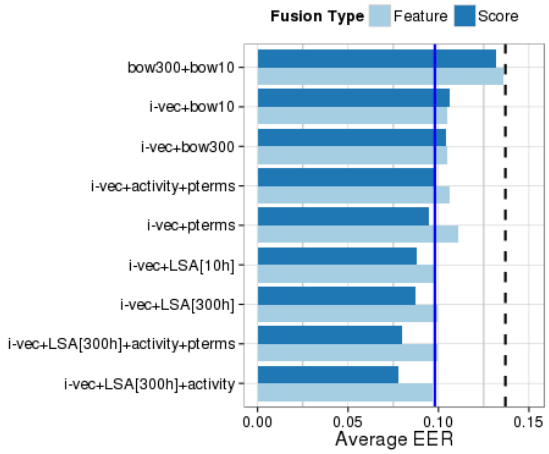

Fig. 2. Score and feature fusion results. BOW baseline is indicated by dashed line, i-vector single system result by solid line.

Table 3. Selected hybrid fusion results (in average EER (%)) including best score fusion for comparison.

Features	Dim.	EER
S(i-vec) + F(LSA[bow300h])	501	9.5
F(i-vec) + S(LSA[bow300h])	601	9.1
F(i-vec) + S(LSA[bow300h]) + S(Activity)	602	8.9
S(i-vec) + S(LSA[bow300h]) + S(Activity)	3	7.8

individual feature result. See the discussion (Sec. 5) for possible explanation. Hybrid fusion results are given in Table 3. While in theory hybrid fusion is more powerful, we were unable to realize improvements over score fusion given our training corpus sizes. This outcome would likely change given a larger training corpus.

5. DISCUSSION

Collection heterogeneity is evident by breaking out performance on a per-collection basis (cf. Figure 3). The top portion of Figure 3 shows the collections for which acoustic i-vectors outperform bags-of-words, and the performance gap between the two features. The bottom portion of Figure 3 shows the same for which bags-of-words performed best. We also indicate the fusion result for each collection with a dot. Although i-vectors were the best overall single feature on average, the bags-of-words (300hr) outperformed i-vectors by at least 2% absolute on 6 of the 22 collections. Intuitively, the collections for which only a single speaker was present (2, 3, 8, 9, 13) all fared better using the i-vector feature (cf. Figure 3). For collections 2, 3, and 13, the difference between i-vector peformance and bag-of-words was at least 10% absolute. Likewise, collections whose primary language was not English (4, 11, 22) fared better with i-vector features. In these cases the notion of relevance is better captured by speaker or language characteristics rather than topic. Conversely, the multi-speaker tutorial collections (5, 16) are more topically coherent and thus better modeled by traditional bag-of-words topic classifiers.

A closer analysis of the fusion results suggests heterogeneity within collections as well. We performed a simple oracle experiment by computing the average over the best single feature on each test collection (i-vec, activity, or LSA[300h]) and measured an 8.8% average EER, whole point worse than the fair score fusion result. This implies that individual collection relevance spans both lexical and non-lexical information. In the majority of cases, fusion is outperforming single features, not just picking the best feature.

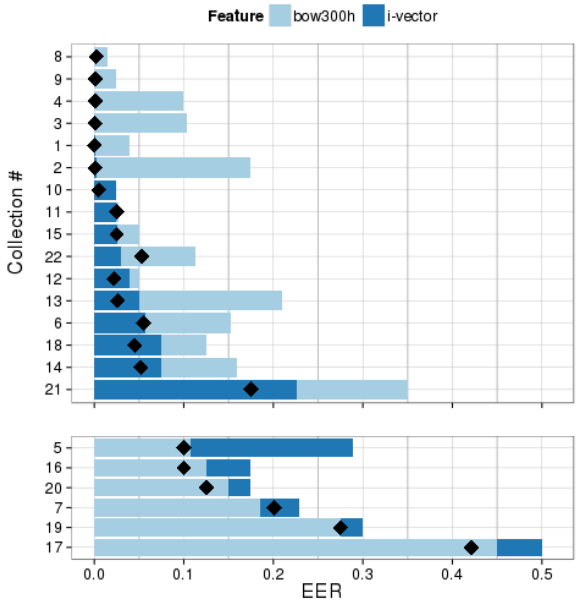

Fig. 3. Impact of lexical, non-lexical and fusion on individual collections, ordered by best performing feature. Bar endpoints indicate individual feature EER for i-vectors and 300h bags-of-words, with the dot indicating the performance from score fusion of the two.

Our most curious findings regard the best performing single feature: the bags-of-senones. These achieve nearly the same error as the fusion of bags-of-words, i-vectors, and speech/music activity, yet the fusion of all four features does not reduce the errors any further. As this is a new task and the information content of bags-of-senones is not yet well-understood, it is impossible to account for this with certainty. However, one explanation is that the bags-of-senones contain signal for several aspects of the multisource fusion that are relevant to this task. This is an intriguing possibility, but proving it would require further exploration into the performance of bags-of-senones on topic ID, speaker recognition, and other potential information sources, in addition to the known value for language recognition.

6. CONCLUSIONS

Content-based recommender systems for spoken documents is an information retrieval task that cuts across traditional speech processing areas such as topic and speaker identification. Using a corpus of internet audio that highlights the multi-modal aspect of a the recommendation task, we evaluated several feature types characterizing all aspects of the content. The system fusion experiments highlight the diverse realizations of what constitutes relevance across users. By fusing lexical and non-lexical-based systems we can reduce the system error by 43% relative (13.7 to 7.8%) to what we would obtain by treating the task simply as topic identification.

7. REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 6, pp. 734–749, 2005.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[3] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen, "Collaborative filtering recommender systems," in *The adaptive web*, pp. 291–324. Springer, 2007.

[4] Yehuda Koren, Robert Bell, and Chris Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[5] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al., "The youtube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 293–296.

[6] Greg Linden, Brent Smith, and Jeremy York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, vol. 7, no. 1, pp. 76–80, 2003.

[7] Michael J Pazzani and Daniel Billsus, "Content-based recommendation systems," in *The adaptive web*, pp. 325–341. Springer, 2007.

[8] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*, pp. 73–105. Springer, 2011.

[9] Raymond J Mooney and Loriene Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the fifth ACM conference on Digital libraries*. ACM, 2000, pp. 195–204.

[10] J-J Aucouturier, François Pachet, and Mark Sandler, """ the way it sounds": timbre models for analysis and retrieval of music signals," *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1028–1035, 2005.

[11] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno, "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences.," in *Proc. ISMIR*, 2006.

[12] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder, "Features for content-based audio retrieval," *Advances in computers*, vol. 78, pp. 71–150, 2010.

[13] John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees, "The trec spoken document retrieval track: A success story," *NIST SPECIAL PUBLICATION SP*, , no. 246, pp. 107–130, 2000.

[14] Barbara Peskin et al., "Improvements in Switchboard recognition and topic identification," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference-Volume 01*. IEEE Computer Society, 1996, pp. 303–306.

[15] Timothy J Hazen, Fred Richardson, and Anna Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 659–664.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[17] Gregory Sell and Pascal Clark, "Music tonality features for speech/music discrimination," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2489–2493.

[18] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011.

[19] John HL Hansen, Rongqing Huang, Bowen Zhou, Michael Seadle, John R Deller Jr, Aparna R Gurijala, Mikko Kurimo, and Pongtep Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 712–730, 2005.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Yanmin Motlicek, Petr a nd Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[21] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.

[22] Luciana Ferrer, Yun Lei, Mitchell McLaren, and Nicolas Scheffer, "Spoken Language Recognition Based on Senone Posteriors," in *Interspeech*, 2014.

[23] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011, pp. 249–252.

[24] David C. Smith, Jeffrey Townsend, Douglas J. Nelson, and Dan Richman, "A Multivariate Speech Activity Detector Based on the Syllable Rate," in *Acoustics, Speech and Signal Processing (ICASSP), 1999 IEEE International Conference on*. IEEE, 1999.

[25] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[26] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.