# MINIMAL-RESOURCE PHONETIC LANGUAGE MODELS TO SUMMARIZE UNTRANSCRIBED SPEECH

*Nancy F. Chen, Bin Ma, Haizhou Li*

Institute for Infocomm Research, A*STAR, Singapore

{nfychen, mabin, hli}@i2r.a-star.edu.sg

## ABSTRACT

We propose to extract summary sentences from lexically untranscribed speech via phone tokenization. We use *decoded phone sequences* instead of *words* to train language models to infer semantically significant utterances. Phone tokens yield comparable results to words on the TDT-2 English corpus, yet require significantly less linguistic resources – no need for automatic speech recognition (ASR): (1) Using decoded phones of high phone error rate (78.7%) leads to comparable results to using ASR-decoded words. (2) Tokenizing English audio using a Czech phone recognizer leads to comparable results to using English words from closed-captions. These trends parallel those established in spoken language recognition and have practical significance: we can potentially summarize speech passages of resource-poor languages by leveraging existing tools developed on resource-rich languages.

***Index Terms—*** extractive speech summarization, phone recognition, phone tokenization, spoken language understanding, spoken document retrieval, audio indexing

## 1. INTRODUCTION

It is more difficult for a user to efficiently browse through audio passages than text passages to determine which particular passages to focus on [1]. This issue could be resolved if an automatic system can provide the most important sentences of the audio passages [2], which is the goal of extractive speech summarization.

Approaches to extract speech summaries include supervised and unsupervised ones. The former is a binary (i.e., summary vs. non-summary) sentence-classification problem. Supervised approaches thus easily take advantage of acoustic, lexical, or structural features [3]. Limitations of supervised methods include imbalanced classes, independence (bag-of-sentences) assumption, and the cost of human-labeled summary sentences. Some researchers have worked on addressing these limitations; e.g., [4] used structured support vector machines to select important utterances while minimizing redundancy, [5] used regression and sampling to make the classes more balanced and [6] used active learning to alleviate the cost of manual labeling.

By contrast, unsupervised approaches do not require human-labeled summaries during training and often model the relationship among sentences (e.g., [7], [8]). Unsupervised approaches often follow the traditional *transcribe-and-summarize* framework: speech is first passed to an automatic speech recognition (ASR) system; the decoded words are then input to a text summarizer. This paradigm takes advantage of techniques from the text domain [9, 10], but also brings forth several challenges [11]. First, ASR is often unavailable, except for resource-rich languages (e.g., English). Second, even if ASR were available for any language, the word transcripts generated by the state-of-the-art systems are still error-prone and suffer from out-of-vocabulary issues. To resolve these issues, past research has focused on acoustic patterns (e.g., [12]) to extract summaries.

In this work, we exploit phonetic patterns instead. We use language models trained on decoded phones to extract summaries, inspired by well-established techniques in spoken language recognition [13]. In addition to language recognition, such phonetic tokenization approaches have been applied in other areas, including spoken term detection [14], spoken document retrieval (e.g., [15], [16]) and topic identification [17]. However, to the best of our knowledge, no study has examined how well phonetic tokenization works on speech summarization. In this work, we attempt to fill in this gap. We want to find out if knowledge in spoken language recognition can carry over to speech summarization in the following two aspects.

(1) When extracting speech summaries, is it necessary to tokenize the audio in lexical units (e.g., words)? Can we relax lexical constraints (e.g., disregard word boundaries), but still summarize well? If so, we can drastically reduce the linguistic resources needed to train unsupervised summarizers: we no longer need ASR - we only need a phone recognizer. (2) Does it matter what language the phone recognizer was trained on? Can we use phone recognizers trained on foreign languages to tokenize spoken documents, yet still achieve robust summarization results? For example, can we summarize English audio using a Czech phone recognizer? If the answers are yes, we can further reduce the linguistic resources to automatically summarize speech passages. To summarize English passages, we do not even need an English phone recognizer - a phone recognizer of *any* language will do. We attempt to address these questions in this paper.

## 2. PHONETIC LANGUAGE MODEL SUMMARIZER

### 2.1. Extractive Summarization Scheme

We adopt the widely-used unsupervised summarization scheme, Maximal Marginal Relevance (MMR) [18]. Given a spoken document $D$, the MMR score of sentence $S_i$ is

$$\text{MMR}(S_i) = R(S_i) - \lambda \max_{j \in \text{summary}} \text{Sim}(S_i, S_j), \qquad (1)$$

where the first term $R(S_i)$ is the relevance function determining how relevant sentence $S_i$ is to the entire document $D$, and the second term is the redundancy penalty for selecting sentence $S_i$ to be in the summary set, which already contains $S_j$. Sentences are iteratively chosen to maximize the total MMR score until the length constraint of the summary is reached. The relevance function $R(S_i)$ is heuristically defined; $\lambda$ is tuned on a development set.

**Table 1**. *Phone recognition systems. ANN: artificial neural network; HMM: hidden Markov model; LM: language model; MFCC: mel-frequency cepstral coefficient; PER: phone error rate; PLP: perceptual linear prediction; STC: split temporal context [24].*

| System | Language | Decoding condition | Corpus | Model | Features | Training data (hr) | # of phones | PER (%) on TDT-2 | PER (%) on original corpus |
|---|---|---|---|---|---|---|---|---|---|
| EN-FA | English | Force Align | WSJ0 [29] | HMM | PLP | 14 | 40 | 0 | 0 |
| EN-BG | English | Bigram LM | WSJ0 | HMM | PLP | 14 | 40 | 52.2 | 19.8 |
| EN-PL | English | Phone Loop | WSJ0 | HMM | PLP | 14 | 40 | 63.5 | 21.7 |
| EN-PL-2 [23] | English | Phone Loop | TIMIT [27] | HMM/ANN | MFCC, STC | 2.8 | 39 | 78.7 | 24.24 |
| CZ [23] | Czech | Phone Loop | SpeechDat(E) [23] | HMM/ANN | MFCC, STC | 9.72 | 46 | N/A | 24.24 |
| HU [23] | Hungarian | Phone Loop | SpeechDat(E) | HMM/ANN | MFCC, STC | 7.86 | 62 | N/A | 33.32 |
| RU [23] | Russian | Phone Loop | SpeechDat(E) | HMM/ANN | MFCC, STC | 14.02 | 53 | N/A | 39.27 |

### 2.2. Phonetic Language Modeling

In this work, we define the relevance function $R(S_i)$ in Eq. (1) as the posterior probability of document $D$ given sentence $S_i$, i.e., $P(D|S_i)$. We take the language modeling approach, where each sentence $S_i$ can be viewed as a probabilistic model for predicting the entire document $D$ [19]. By further assuming that speech units (e.g., words) are conditionally independent given the sentence $S$, and their order is of no importance, $P(D|S)$ can be decomposed into a product of unigram probabilities of speech units $u$ generated by sentence $S$. The relevance score of sentence S is therefore defined as:

$$R(S) = P(D|S) = \prod_{u \in D} P(u|S)^{c(u,D)}, \tag{2}$$

where $P(u|S)$ is the unigram probability of speech unit $u$ given the sentence $S$, which can be obtained through maximum likelihood estimation; $c(u, D)$ is the number of times that speech unit $u$ occurs in document $D$. In this work, we propose to extend the speech units $u$ from words to *n-grams of decoded phones (from any language)*.

The unigram probabilities in Eq.(2) were determined through maximum likelihood estimation:

$$P(u|S) = \frac{c(u, S)}{|S|}, \tag{3}$$

where the speech unit $u$ are words, $c(u, S)$ is the number of times word $u$ occurs in sentence $S$, and $|S|$ is the number of words in $S$.

The redundancy penalty in the second term of Eq.(1) was based on vector space modeling using cosine similarity. Each sentence $S_i$ is represented in vector form, where each dimension $z_{t,i}$ specifies the product of the term frequency (TF) and inverse document frequency (IDF) of the word $u_t$. The conventional cosine similarity between sentence $S_i$ and $S_j$ is then computed as:

$$\text{Sim}(S_i, S_j) = \frac{\sum_{t=1}^{V} z_{t,i} \times z_{t,j}}{\sqrt{\sum_{t=1}^{V} z_{t,i}^2}, \sqrt{\sum_{t=1}^{V}, z_{t,j}^2}} \tag{4}$$

where $V$ is the number of distinct words.

## 3. EXPERIMENTS

### 3.1. Setup

#### *3.1.1. TDT-2 English Corpus*

TDT-2 [21] was used in the NIST Topic Detection and Tracking evaluations [20]. We used 114 ABC broadcasts of 1,357 English news stories (43 hrs) partitioned randomly into the development and test set. The reference summaries and utterance boundaries are the same as in [22][1]. The length constraint of the extracted summary was set at 5% of each story. We only evaluated short summaries since each reference summary from [22] only contains one sentence.

#### *3.1.2. Phone Recognition Systems*

We adopted state-of-the-art phone recognizers [23, 24, 25] developed at Brno University of Technology (BUT). These phone recognizers were successfully applied to the 2005 NIST language recognition evaluation [26]. The non-English (Czech, Hungarian, Russian) phone recognizers were trained on SpeechDat(E) [23], while the English one was trained on TIMIT [2] [27]. Main attributes of these phone recognizers (denoted as CZ, RU, HU, EN-PL-2) are listed in Table 1; see [23] for more details.

Since TIMIT is relatively small in size, we also trained an English phone recognizer on WSJ0 [29] (3states/phone; 32 Gaussians/state; state-clustered triphones). As shown in Table 1, the WSJ0 recognizer decoded phones under three conditions: (1) force alignment (EN-FA) using closed caption transcripts; (2) phonetic bigram language model (EN-BG), where the language model was trained on WSJ0; (3) phone loop (EN-PL).

For the English systems, phone error rate (PER) was measured on TDT-2 and their original corpus (WSJ0 or TIMIT). The phones obtained through forced-alignment (EN-FA) were used as ground-truth for PER computation. For the non-English BUT phone recognizers, we only included the PER obtained from the SpeechDat(E) corpus, since it is meaningless to measure PER on TDT-2 if the reference phones are in English and the decoded phones are non-English.

#### *3.1.3. Evaluation Metrics*

We adopted the widely-used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [30]; in particular, we used ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-N is an n-gram co-occurrence statistics counting the number of overlapping n-grams of words between the automatic summary and the reference summary:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} \text{Count}(gram_n)}$$

where $n$ is the length of the n-gram, $gram_n$; $\text{Count}_{\text{match}}(gram_n)$ is the maximum number of n-grams co-occurring in an automatic

[1] We thank Heidi Christensen for sharing human determined extractive summaries and utterance boundaries with us.

[2] Phone recognition performance comparable to that of [28] (PER=24%), trained by deep belief networks.

summary and the reference summary. When $n = 1$, the metric is ROUGE-1. ROUGE-L measures the longest common subsequence.

We assumed the extracted summary is presented in audio (instead of text) as is in [19]; i.e., when using decoded words from ASR, decoded words were used to compute scores using Eq.(2), but once a candidate summary was chosen, the ROUGE evaluation uses the words from the closed caption transcripts (since the output audio does not suffer from ASR errors) to evaluate the automatic summary.

We also included a stricter measure than ROUGE, *accuracy*. Accuracy is 1 if the extracted summary of a spoken document *exactly matches* its reference summary; i.e., accuracy is 0, even if only one word in the extracted summary differs from the reference summary.

### 3.2. Experimental Design

Language models expressed in Eq.(2) are computed using three types of speech units $u$: (1) English word tokens, (2) English phone tokens, (3) foreign (non-English) phone tokens.

#### *3.2.1. English Word Tokens*

We used two sets of English word tokens from TDT-2 English [21]: (1) closed caption transcripts; (2) decoded words from the Dragon ASR system (word error rate=30%).

#### *3.2.2. English Phone Tokens*

We decoded the spoken documents using the four English phone decoding conditions (EN-FA, EN-BG, EN-PL, EN-PL-2) in Table 1. We used 3-grams (i.e., 3 consecutively decoded phones) as the speech unit $u$ in Eq. (1). This modeling choice was determined empirically on the development set (3-grams performed slightly better than 2-grams and 4-grams) and consulting prior work in topic identification [17] and spoken document retrieval [16].

#### *3.2.3. Foreign Phone Tokens*

We tokenize the English audio using the three non-English (Czech, Hungarian, Russian) phone decoding conditions (CZ, HU, RU) in Table 1.

## 4. RESULTS AND DISCUSSION

Summarization results vs. phone error rate (PER) for ROUGE-1, ROUGE-L, and accuracy are shown in Fig. 1; all show similar trends. (Results for ROUGE-2 are not shown due to space constraints, but ROUGE-2's trends are the same as the other three metrics.) For English phone decoding, we show phone error rates on the TDT-2 corpus (top panel) and those on the original corpus used to train the phone recognizer (bottom panel). For non-English phone decoding, only phone error rate on the original corpus are shown (bottom panel) since phone error rate cannot be measured if the reference phones are English and the decoded phones are non-English.

Summarization results using English word tokens from closed caption and ASR are statistically similar except for ROUGE-1, where $p \approx 0.05$. These results are baselines for the following experiments using phone tokens.

### 4.1. English Phones Summarize as Well as English Words

We observe the following trends when comparing English word tokens and English phone tokens in Fig. 1: (1) Summarization performance is inversely correlated with phone error rate (PER). (2) Summarization results using English phone tokens from system EN-FA (PER=0) is always better than English word tokens, be them from closed caption or ASR ($p < 0.05$). (3) When PER (on TDT-2 corpus) is above 50% (e.g., EN-BG, EN-PL), phone tokens achieve similar summarization results as baseline word tokens.

#### *4.1.1. Summarization Results Correlate with Phone Accuracy*

The absolute value of Pearson correlation coefficient $|r| \geq 0.95$ for all blue curves with diamond markers; see Fig. 1 for details. When PER on TDT-2 corpus increases from 0% to 52.2%, 63.5%, and 78.7% absolute, the relative drop in performance (averaged across all four metrics) are 6.3%, 8.7%, 11.8% , respectively.

#### *4.1.2. High Phone Error Still Yields Comparable Results to Words*

While low PER leads to better summarization results, high PER still compares favorably with those using English word tokens: When PER is 52.2% (EN-BG) and 63.5% (EN-PL), phone tokens achieve similar results as word tokens (closed-caption). Even when PER is 78.7% (EN-PL-2), phone tokens achieve similar results as ASR word tokens. These results imply that while the decoded phones are not the true phonetic identities, they can still be used to infer semantics to extract reasonable summaries. Our results suggest that phonetic language models are cost-effective for extracting speech summaries, as the linguistic resources required to train a phone recognizer are much lower than those for an ASR system.

#### *4.1.3. Why Phone Tokenization Works*

Our results imply that semantic significance of a spoken document can still be inferred when lexical constraints are relaxed: (a) word boundary relaxation (e.g., *recognize speech* and *wreck a nice beach* are almost phonetically identical); (b) homophone relaxation: distinct words that sound the same but have different meanings are no longer differentiated (e.g., *no* and *know* both become [n ow]).

Word boundary relaxation might not affect our summarization model much, since word boundaries can be implicitly inferred from co-occurrences of n-grams of phones that make up a word. For example, if the word *recognize* is topic-revealing, trigrams of phones modeling the word *recognize* will co-occur frequently in many sentences. The sentences containing the consecutive words *wreck a nice* would likely be significantly fewer and thus act as noise, so *wreck a nice* would not overshadow the true keyword *recognize*.

We conjecture homophone relaxation is not a major issue either, since few words sound the same but have different meanings in English, where the ratio of distinct words vs. different sounding words is 3.2 (computed from [31]). At first glance, homophone relaxation could present challenges in languages like Mandarin, where the ratio of distinct characters vs. different sounding characters is 25 if disregarding tones and 7.7 if considering tones (computed from [32], [33]). But homophone relaxation might not affect Mandarin speech summarization severely: Similar sounding characters are often disambiguated by its contextual characters; i.e., while many characters sound the same, words are often formed with two or more characters, and are thus not as acoustically confusable. Therefore, Mandarin words can still be modeled by co-occurring trigrams of phones as described in word boundary relaxation. How other languages like Mandarin are affected by these relaxations are open research questions worthy of future investigation.

#### *4.1.4. Perfect Phone Tokens Outperforms Perfect Word Tokens*

It is intriguing that PER=0 (EN-FA) performs better than using word tokens. We suspect this improvement stems from distinguishing spoken words identical in spelling but different in pronunciation. This

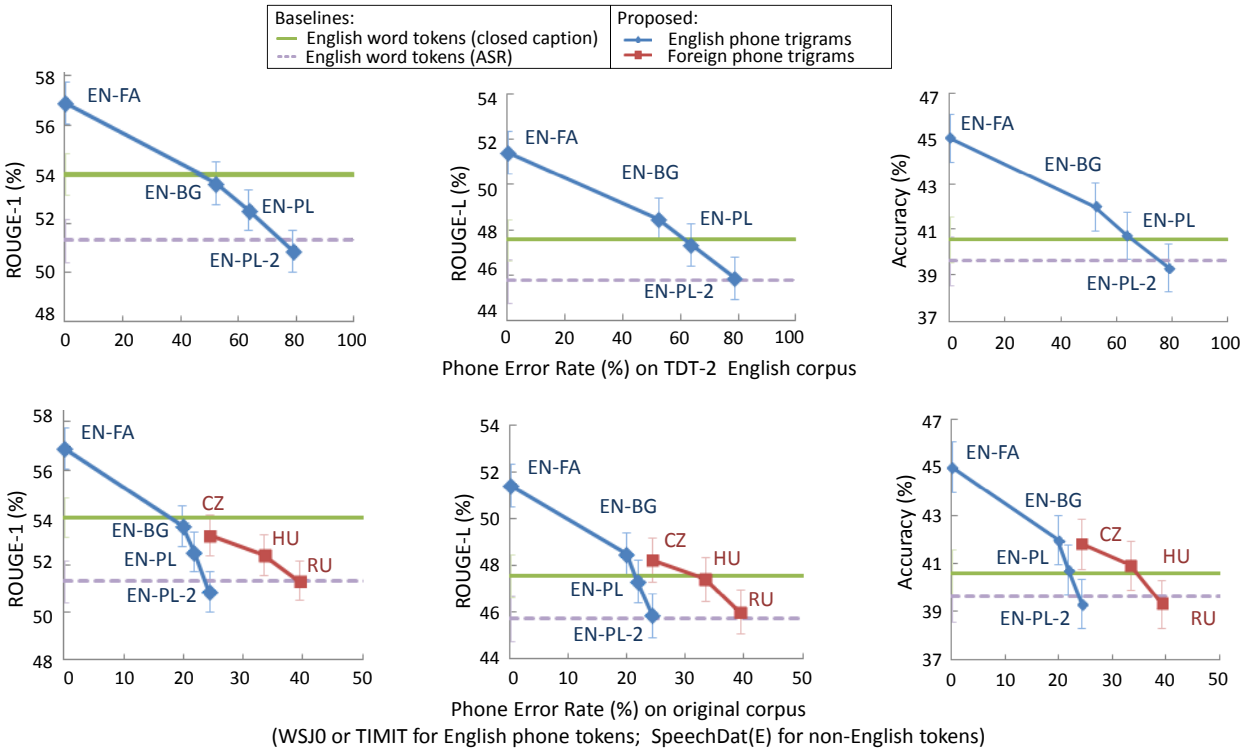


**Fig. 1**. *Performance and phone error rate (PER) are inversely correlated: average correlation coefficient for English phones (blue-diamond curves) are* $-0.99$ *(top panel),* $-0.95$ *(bottom pannel) and that for non-English phones (red-square curves) is* $-0.96$*. Summarization results for non-English phones not applicable for top panel (PER on the TDT-2 English corpus) since one cannot compute PER when the language of the decoded phone set (e.g., Czech) differs from that of the reference phone set (English). Error bars represent standard error. Refer to Table 1 for details of the phone recognition systems EN-FA, EN-BG, EN-PL, EN-PL-2, CZ, HU, RU.*

difference potentially correlates with whether the spoken word is in a summary sentence or not. These differences arise from speaking styles (e.g., important words spoken with less reduction) or dialect variations. For example, news anchors usually speak in the mainstream dialect. Summary sentences tend to be from them instead of interviewees, who more likely speak in non-mainstream dialects. Assume *dark knight* is a key term of a news story. While a Bostonian interviewee would likely delete the /r/ in *dark*, the news anchor would not. The key term is modeled as [d aa r k n ay t] instead of [d aa k n ay t], though both refer to the same lexical identity. We plan to further investigate the factor of speaking style in future work.

### 4.2. Foreign Phone Tokens

#### *4.2.1. Foreign Phones Summarize as well as English Words*

Fig. 1 shows that English phone tokens do not always perform better than foreign phone tokens. Summarization results from CZ are better than those from EN-PL and EN-PL-2 and comparable to EN-BG and English word tokens (closed-caption). Similarly, summarization results from HU and RU are comparable to EN-PL and EN-PL-2, respectively. From Fig. 1, we see that for foreign phone tokens (CZ, HU, RU), summarization results are also inversely correlated ($r = -0.96$) with phone error rate (on the SpeechDat(E) corpus). These correlations parallel findings in language recognition [26].

These results suggest that when summarizing English audio, we neither need an English ASR system, nor an English phone recognizer – all we need is a phone recognizer trained on *any* language. Our findings suggest that we can significantly reduce linguistic resources by leveraging existing tools developed on resource-rich languages (e.g., English) to automatically summarize spoken documents in languages like Min Nan (aka Hokkien, Taiwanese; at least 47 million speakers [34]) or Malay ($\approx$ 210 million speakers [35]).

#### *4.2.2. Why Foreign Phone Tokenization Works*

Similar to language recognition, tokenizing speech to phones of foreign languages characterizes the underlying acoustics. The higher the accuracy of the phone recognizer, the better the characterization power. This characterization helps infer semantically significant portions of the audio. It does not matter which phone labels (be them Czech or English) are used to map and model the underlying semantic units, as long as this mapping is consistent. Our phonetic tokenization framework models similar information as acoustic approaches [12], but we can directly leverage existing high performance phone recognizers used in ASR or language recognition [23]. This leverage drastically reduces the required training time.

## 5. CONCLUSIONS

We proposed to extract summaries from lexically untranscribed speech via phonetic language models; instead of using words to infer semantic significance, we used decoded phones. We showed that established knowledge in language recognition can be carried over to speech summarization on the TDT-2 English corpus: (1) The higher the accuracy of the phone recognizer, the better the summarization results. (2) The language of the phone recognizer need not be English to summarize English audio. Czech phone tokens can lead to better summarization results than English phone tokens, given sufficiently low phone error rate. (3) High phone error rate (78.7%) still leads to summarization results similar to English word tokens decoded from the Dragon ASR system. These findings suggest that we can summarize speech of any language even with limited resources – all we need is a phone recognizer of *any* language; no ASR is needed. For future work, we plan to investigate whether these findings are generalizable to speaking styles like meetings.

## 6. REFERENCES


[1] Kong, S.-Y. and Lee, L.-S., "Semantic Analysis and Organization of Spoken Documents Based on Parameters Derived From Latent Topics", IEEE Trans. Speech and Audio Proc., 19(7):1875–1889, 2011.

[2] Christensen, H., Gotoh, Y., and Renals, S., "A cascaded broadcast news highlighter", IEEE Trans. Audio, Speech and Language Proc., 16.1: 151–161, 2008.

[3] Maskey, S. and Hirschberg, J., "Comparing Lexical, acoustic/prosodic, discourse, and structural features for speech summarization", Proceedings of Eurospeech, 2006.

[4] Lee, H-Y., Chou, Y-Y., Wang, Y-B., Lee, L-S., "Supervised Spoken Document Summarization Jointly Considering Utterance Importance and Redundancy by Structured Support Vector Machine", Proceedings of Interspeech, 2012.

[5] Xie, S. and Liu, Y., "Improving Supervised Learning for Meeting Summarization using Sampling and Regression", Computer Speech & Language, 24.3:495-514, 2010.

[6] Zhang, J. and Fung, P. "Active learning with semi-automatic annotation for extractive speech summarization." ACM Transactions on Speech and Language Processing, 8.4: 6, 2012.

[7] Garg, N., Favre, B., Reidhammer, K., and Hakkani-Tr, D., "ClusterRank: a graph based method for meeting summarization," Proceedings of Interspeech, 2009.

[8] Chen, Y-N. and Metze F., "Integrating Intra-Speaker Topic Modeling and Temporal-Based Inter-Speaker Topic Modeling in Random Walk for Improved Multi-Party Meeting Summarization", Proceedings of Interspeech, 2012.

[9] Erkan, G. and Radev, D. R., "LexRank: Graph-based lexical centrality as salience in text summarization", J. Artif. Intell. Res. 22 457–479,2004.

[10] Mihalcea, R. and Tarau, P., "TextRank: Bringing order into texts", Proceedings of EMNLP. Vol. 4. 2004.

[11] Furui, S., "Recent Advances in Automatic Speech Summarization", Proceedings of IEEE Spoken Language Technology Workshop, 2006.

[12] Zhu, X., Penn, G., and Rudzicz, F. "Summarizing Multiple Spoken Documents: Finding Evidence from Untranscribed Audio", 47th Annual Meeting of the Association for Computational Linguistics (ACL), 2009.

[13] Zissman, M., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE TSAP., 4(1):31-44, 1996.

[14] James D. A. and Young, S. J., "A fast lattice-based approach to vocabulary independent wordspoting", Proceedings of IEEE ICASSP, 377–380, 1994.

[15] Shen, W., White, C. M., and Hazen, T. J., "A Comparison of Query-by-Example Methods for Spoken Term Detection", Proceedings of Interspeech, 2009.

[16] Chia, T. K., Sim, K. C., Li, H., and Ng, H. T., "Statistical lattice-based spoken document retrieval", ACM Transactions on Information Systems 28(1), 2, 2010.

[17] Hazen, T. J., Richardson, F., and Margolis, A., "Topic Identification from Audio Recordings using Word and Phone Recognition Lattices," Proceedings of IEEE ASRU, 2007.

[18] Carbonell, J., Goldstein, J. "The use of MMR, diversity-based reranking for reordering documents and producing summaries," Proc. of ACM SIGIR on Research and development in information retrieval, 1998.

[19] Chen, B. and Lin, S.-H., "A Risk-Aware Modeling Framework for Speech Summarization", IEEE Trans. Speech and Audio Proc., 20(1):199–210, 2012.

[20] National Institute of Standards and Technology (NIST) Topic Detection and Tracking (TDT) Evaluation: http://www.itl.nist.gov/iad/mig/tests/tdt, last accessed 16 November, 2012.

[21] Linguistic Data Consortium, "TDT-2 Corpus: http://projects.ldc.upenn.edu/TDT2/", University of Pennsylvania, last accessed 16 November, 2012.

[22] Christensen, H., Kolluru, B., Gotoh, Y., and Renals, S., "From Text Summarisation to Style-Specific Summarisation for Broadcast News", European Conference on Information Retrieval, 2004.

[23] Schwarz, P., Matejka, P., Burget, L., and Glembek, O., "Phoneme recognizer based on long temporal context", Brno University of Technology. Online: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context, accessed on 25 Oct, 2012.

[24] Schwarz, P., "Phoneme Recognition based on Long Temporal Context", PhD Thesis Brno University of Technology, 2009.

[25] Schwarz, P, Matejka, P., Cernocky, J., "Hierarchical Structures of Neural Networks for Phoneme Recognition", Proc. of ICASSP, 2006.

[26] Matejka, P. et al, "Phonotactic Language Identification using High Quality Phoneme Recognition", Proceedings of Eurospeech, 2005.

[27] Garofolo, J. S., "TIMIT: Acoustic-phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.

[28] Mohamed, A. R., Dahl, G., and Hinton, G., "Deep belief networks for phone recognition," In NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

[29] Paul, D. and Baker, J. "The design for the Wall Street Journal-based CSR corpus," Proceedings of DARPA Speech and Natural Language Workshop, Morgan Kaufmann, 1992.

[30] Lin, C.-Y., "ROUGE: A Package for Automatic evaluation of summaries", Workshop Text Summarization Branches Out, 2004.

[31] English homophones: http://www.homophone.com/, last accessed 16 November, 2012.

[32] The Dictionary of Variant Forms of Chinese Characters: http://dict.variants.moe.edu.tw/start.htm, last accessed 16 November, 2012.

[33] Lee, C.-H., Li, H., Lee, L.-S., Wang, R., and Huo, Q. (editors), "Advances in Chinese Spoken Language Processing," World Scientific, 2007.

[34] Nationalencyklopedin (Swedish language encyclopedia), "Vrldens 100 strsta sprk 2007 (The World's 100 Largest Languages in 2007)," 2007.

[35] Uli, K., "How many people speak Indonesian," University of Hawaii at Manoa, retrieved 20 October 2012.