INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Targeted Keyword Filtering for Accelerated Spoken Topic Identification

Jonathan Wintrode

Raytheon Applied Signal Technology, USA
jonathan.c.wintrode@raytheon.com

Abstract

We present a novel framework for spoken topic identification that simultaneously learns both topic-specific keywords and acoustic keyword filters from only document-level topic labels. At inference time, only audio segments likely to contain topic-salient keywords are fully decoded, reducing the system's overall computation cost. We show that this filtering allows for effective topic classification while decoding only 50% of ASR output word lattices, and achieves error rates within 1.2% and precision within 2.6% of an unfiltered baseline system.
Index Terms: speech recognition, keyword spotting, topic identification

1. Introduction

We find instances of spoken topic identification (ID) under various forms across the spectrum of speech processing tasks. From providing a discrete categorization for call center routing [1], recommender systems for speech content [2], theme identification for improving dialog systems, [3], task identification for personal assistants (ref), and spoken document retrieval (SDR), many tasks and systems benefit from a coarse representation of information derived from content-bearing words.

The topic content of spoken documents, found in Automatic Speech Recognition (ASR) output, tends to concentrate in a relatively small number of word tokens and are often repeated in a local context. This redundancy of information is consistent with a computational linguistic analyses of 'burstiness' in text corpora the underlies traditional information retrieval statistics like document frequency (cf. [4, 5]). We can confirm our intuition by artificially increasing the error rate of ASR systems while attempting to correctly recognize specific keywords - without significantly impacting topic classification performance (cf. [6]).

Here we extend these conclusions to speed up the ASR pipeline for spoken topic ID by fully decoding only topically relevant segments of audio. We simultaneously learn salient keyword features for the topic classification task and build acoustic keyword detectors to avoid lattice generation for a large subset of the audio data during inference. Our initial keyword (KW) detection work [7] required user annotation of KW occurrences at the segment level. By learning detectors (used as filters) in the context of topic ID, we reduce the annotation burden from 500-10000 KW segment labels to 10-100 document-level labels per topic.

Given that error-robustness of topic information is not a language-specific phenomenon, we want to demonstrate that a keyword-filtered topic ID pipeline (as illustrated in Figure 1) is effective in multiple languages, not just English. Using spoken topic ID tasks in both English and Mandarin we can in fact show that document-level annotations are sufficient to train topically relevant keyword filters. We can filter out - avoid fully decoding - over half of the test audio with only minimal degradation in overall topic classification performance. For the English task

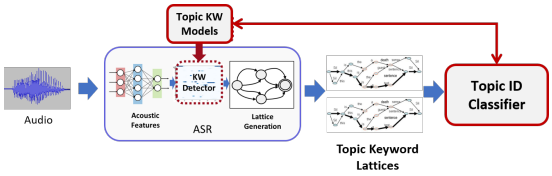

Figure 1: *Pipeline for paired spoken Topic ID with KW filtering.*

we observe a reduction in average precision of only 1.2% absolute and less than 1% increase in detection error rate (EER). For Mandarin, with a 50% reduction in fully decoded segments, we see a 2.6% reduction in average precision and increases of less than 2% in the EER. If our application can handle more errors we can reduce full decoding up to 80% while increasing the topic detection EER by 5% or less.

2. Related Work

Since early work on AT&T's "How may I help you" [1], poken topic ID has relied on the classification of word-based ASR output to produce bag-of-word type features. These can then be used as input to any and all text-based supervised classification methods. The community has explored variations on this pipeline such as phonetic-based recognition (cf. [8]) and other sub-word techniques assumed to be faster and requiring fewer resources for recognition training. However, these non-word approaches in general produce systems with roughly double the classifier error rate.

Other alternative features have been considered, from purely acoustic features such as GMM-based i-vectors[2] to self-discovered or self-organizing units (SOUs) [9, 10]. For the content-based recommender work in [2], the categories were not exclusively determined by semantic content (spoken languages or specific speakers strongly indicated the category). Given this the i-vectors outperformed ASR word input for some classification tasks. However, for the self-discovered phonetic or lexical units, which focused primarily on ASR training constraints, saw similar topic performance degradation as phonetic approaches compared to word-based systems. In relation to this work it is worth noting the SOU approach, in spite of its constraints, did exhibit some amount of keyword discovery correlated to the topic categories.

At the word level, much of the work done on spoken topic ID has naturally focused on different classification techniques drawn from text classification literature (see [11] for an overview of pre-deep learning techniques). More recent neural network and word embedding based techniques (cf. [12, 13, 14]), are of interest to the spoken topic ID task as well generative and unsupervised latent topic modeling approaches.

Approaches to the spoken topic task, typically using either Fisher or Switchboard informal English speech using genera-

http://dx.doi.org/10.21437/Interspeech.2021-1670

tive topic models [15, 16] or similar Bayesian models [17] have achieved comparable but not significantly better accuracy to bag-of-words approaches. From a runtime perspective, though, these approaches still require generation of full word recognition output.

In terms of the ASR systems generating word features, our keyword filtering approach assumes we have frame-level output from a typical HMM-DNN 'hybrid' system exemplified by [18]. With more focus on end-to-end or sequence-to-sequence ASR (cf. [19]) and the growth of transformer-based architectures, the use of these model architectures are naturally of interest to topic-related tasks. BLSTM models have achieved superior results on an utterance level topic spotting task on the Switchboard corpus [20], and the use of transformers for on-device keyword spotting [21] suggests that these types of models could be made lightweight enough for our topic task as well.

3. Methods

Topic ID with keyword filtering encompasses three interrelated learned tasks. First we select a targeted ***feature set of keywords*** $W = w_1, w_2, \ldots, w_K$, given a training set of spoken documents $D = d_i, d_2, \ldots, d_N$ labeled with topics $T = t_1, t_2, \ldots, t_M$. Second, for the keywords in W we build a binary ***keyword set detector*** for filtering topically irrelevant segments at runtime from the training utterances. Third, to match the desired runtime scenario we use the keyword set W and train a $M-$way ***topic classifier*** using only segments in D that contains a keyword from W.

For all tasks, we assume we *do not* have ground truth transcripts for the topic-labeled training corpus D. The learning tasks are assumed to occur offline and as a first step, we generate word lattices for each utterance in the training set D. Expected word counts are then computed from lattice posteriors. A consistent token stream, not perfect keyword recognition accuracy, is sufficient to identify salient keyword features. In previous work, using lattice word posteriors as topic ID bag of words features we showed that performance was quite insensitive to the WER [6]. We can reasonably claim that the methods here identify a topically relevant keyword set filter without requiring perfect accuracy for success.

3.1. Feature Selection and Topic Classification

To select a set of topic-salient keywords for the classification task we apply standard feature selection statistics described in [22]. We apply the χ^2 metric, defined as a binary statistic between two variables (the keyword and topic label) to rank each word observed in the training lattices. We weight token counts by the lattice posterior and select the top K-scoring words as our keyword set.

We considered two keyword selection scenarios: *multi-class labeling*, corresponding to the typical multi-class supervised training scenario where we build a single M-way classifier and select one set of keywords, and *single-topic labeling* where we consider the case where potential users of the system define topics independently in which case we build M binary classifiers and M keyword set filters.

For the single topic case, we have a score per topic per keyword. For the multi-class scenario, we rank keywords by taking the maximum value over all M topics. In contrast to the feature selection approach originally presented in [22], where the goal is to reduce dimension of the classifier input space, we use the keyword list to limit the number training and test segments

Figure 2: *Example coarse phonetic bins from AM output posteriorgram containing topic keywords "minimum" and "wage".*

from which we compute bags of words for topic classification. We considered information gain as well, which in [22] worked as well as χ^2, however we saw worse performance on human transcripts so all results refer to the χ^2 ranking.

Taking all utterances in D that contain one or more of the K keywords we can train a linear classifier. The feature space is still the full ASR vocabulary size. We perform our experiments using the Naive Bayes formulation described in [8]. Our approach does not preclude other classification methods, however for our corpora more complicated methods have not yet shown significant accuracy improvements.

3.2. Keyword Set Detection

Here we briefly describe the keyword filtering techniques developed in [7]. Given acoustic model output probabilities O for utterance s_i in document d_j can we determine if *any* of the keywords in W occur in s_i? We learn a binary classifier with scoring function $S = f(O, W)$ to each segment and fully decode s_i only if S is greater than some threshold t.

We express O as a posteriorgram, a $f \times p$ matrix where f is the number of frames in s_i and p is the number of phonetic units in the model. For the English model we had $p = 46$ and for Mandarin, $p = 67$. Following previous experiments, we aggregated context-dependent phone probabilities from the target ASR system to the center phone of each output state.

We apply the two best-performing classifiers from [7], first, a Naive Bayes classifier using discrete bags-of-phone-n-gram features derived from the posteriorgrams and second, and direct classification of the posteriorgrams using a 3-layer CNN with softmax output. Both proved effective for the keyword filtering task, however as we will show subsequently, the CNN performs consistently better in the context of filtering for topic ID.

To compute discrete 'bags of phones' for linear classifiers, we segment the posteriorgram using a sliding window over the output distributions. We mark putative phone boundaries where the root mean squared error (RMSE) between adjacent windows exceeds some threshold t_c. As before, for all our experiments we use window of 2 frames and set t_c to 0.001. We normalize posteriors of each region between change points to sum to 1 and thus generate expected phone counts. This binning is illustrated in Figure 2, and from this we extract 2-gram sequences for the Naive Bayes classifier training and inference.

For the CNN classifier, whose structure is given in Table 1, we take the phone posteriorgrams as input and output a KW or $NO - KW$ binary decision for the segment. We pad input with SIL phones for a consistent 10s (1000 frame) input. We considered more complex LSTM-based models in [7] but found no additional benefit above sequence models.

Table 1: *Model components for CNN KW detection classifier.*

Component	In	Out
Conv2D + MaxP + Drop	(1000,p)	256
Conv2D + MaxP + Drop	256	256
Conv2D + MaxP + Drop	256	128
Dense Relu	256	64
Dense Sigmoid	64	1

4. Experiment Setup

The guiding hypothesis underlying the efficacy of keyword filtering for spoken topic ID is that topic information tends to be robust and redundant to ASR errors and that this phenomenon extends across languages. To this end we test our methods against topic-annotated corpora in English and Mandarin Chinese. For English, we use the now well-studied Fisher English topic-labeled conversational speech corpus[23] with the topic train and test partitions originally defined in [8].

For Mandarin, we divided the HKUST Mandarin Telephone Speech corpus [24] which was constructed with a similar methodology to the Fisher English corpus such that each conversation had an associated topic prompt used as the ground truth label. The number of recordings per topic varied significantly, so we selected the 15 most frequent topics, ensuring at least 20 recordings per topic. The LDC-provided train/dev split (intended for an ASR task) was highly skewed so we randomly selected approximately 33% of the combined sets as a test set and reserved the rest for training, matching the train/test split for Fisher English as best as possible.

Table 2: *Topic ID corpus sizes and number of topics (M)*

Corpus	Train (N)	Test	M
Fisher English	2,748	1,372	40
HKUST Mandarin	1,000	414	15

The document counts given in Table 2 consider each side of a conversation as an independent document. Previous work has shown that the topic classification task is improved by combining information from both conversation sides and our work does not preclude this benefit. As we focus on injecting topic information to speed the ASR decoding process, which generally operates on one side at a time, we evaluate the topic task one side at a time as well.

Finally, we test our topic-based filtering and identification framework using well-trained but not overly matched off-the-shelf Kaldi-based ASR systems. For English, we trained a standard TDNN-HMM system similar to the one described in [18] however we omitted the Fisher corpus from the recipe to avoid having our topic data in the ASR training set.

For Mandarin, we selected the CVTE TDNN model from kaldi-asr.org [25], which was trained on a private commercial data set. Again we needed to ensure the HKUST data was not used in the ASR training corpus. Thus for both languages we look at the performance under less than perfectly matched domains, which is a condition we would actually expect often in real-world conditions.

As with previous work we look at spoken topic ID, particularly for big data scenarios, as a detection task. We report an average precision metric, the area under the precision-recall curve averaged over each topic label (mAUC). We also report the equal error rate (EER) averaged across topics, whether trained as a multi-class system or for the single topic filter scenario, M independent 1-vs-all binary classifiers.

Table 3: *Performance summary for Fisher English task.*

Condition	K	mAUC	EER
Transcripts - all segments	-	78.1	4.48
Transcripts - filtered	500	77.0	5.02
	1000	77.6	4.69
	2000	77.8	4.37
ASR lattices - filtered	500	75.2	7.08
(exact match)	1000	76.0	6.60
	2000	76.7	6.28
ASR lattices - all segments	-	**76.8**	**6.87**
NB filter @ 50%	500	72.1	8.35
	1000	73.9	8.01
	2000	74.5	7.53
CNN filter @ 50%	500	75.4	7.08
	1000	75.6	7.19
	2000	**75.6**	**7.05**

5. Results

Our experiments focus not just on the overall accuracy and error characteristics of the task but rather the performance of the topic ID system with the associated keyword filters applied. We are interested in performance relative to how much data we don't have to decode. Tables 3 and 4 show a synopsis of our results for each dataset.

As a set of upper bounds, we first show performance of the topic classifier evaluated on either human transcripts or the full set of ASR lattices. Where K is specified, we have filtered out segments from the train and test features where none of the top K keywords occur. For transcripts, keyword occurrence is captured by a string match, for ASR lattices, we say a keyword occurs if the lattice posterior is greater than 0.5. This latter condition is closer to a fair test, but we still have to decode the lattice to see if the keywords occur.

These oracle systems are noted in the tables with italics. For the Fisher English data we see the largest performance gap is actually between the ground truth system and the ASR oracle-filtered lattices. We lose less than 2% in terms of average precision (mAUC) and increase our EER by 2% for matched K conditions when adding noise from the ASR system. For the HKUST data, we lose at most 2% in precision and see comparable EERs for the full ASR systems.

For filtered systems using the NB bag-of-n-grams and CNN KW detectors we expect added error having to detect topically-relevant segments, but our results show that this has a small impact on the final topic classification performance. In Tables 3 and 4 we report the topic ID performance at the threshold which filters out 50% of the total test segments and then look at the whole range of operating points at the end of this section. We examined K values larger than the 2000 shown below but we saw no significant performance difference beyond that.

At the 50% threshold, we lost only 1.2% (absolute) in mAUC on the English set and 2.6% on Mandarin with the top-performing CNN-based KW filters. We increased EER by only 0.2% for the English set and by 1.3% for Mandarin. For En-

Table 4: *Performance summary for HKUST Mandarin task.*

Condition	K	mAUC	EER
Transcripts - all segments	-	97.2	2.17
Transcripts - filtered	500	86.9	7.77
	1000	90.6	5.33
	2000	90.5	5.29
ASR lattices - filtered	500	88.7	6.98
(exact match)	1000	90.6	5.75
	2000	91.0	5.73
ASR lattices - all segments	-	**93.1**	**4.82**
NB filter @ 50%	500	84.8	9.52
	1000	87.9	8.31
	2000	89.9	7.23
CNN filter @ 50%	500	86.6	8.22
	1000	89.0	7.02
	2000	**90.5**	**6.05**

glish, the CNN filtered experiments were less sensitive to the choice of K and in both cases the CNN filters consistently outperformed the NB filters in selecting segments for the topic task.

Figures 3 and 4 show the detection performance across all possible KW filter thresholds. The CNN filters are shown with solid lines and the NB filters with dashed. Up to $K = 2000$, more keywords are more effective (performance plateaued at $K = 3000$ which we omit for brevity). Depending on application requirements, if we only wish to fully decode 25% of segments we still only increase our EER by 2.3 and 2.6% on the English and Mandarin sets respectively. Furthermore, we do not need to annotate all 2000 keywords, rather they are learned

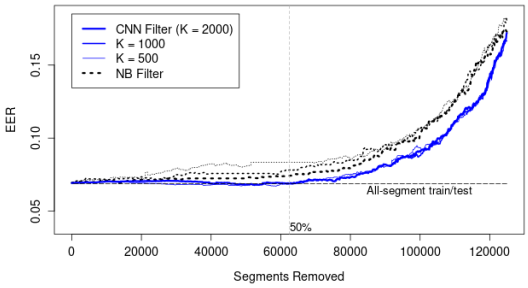

Figure 3: *Effect of KW filters on English topic detection error.*

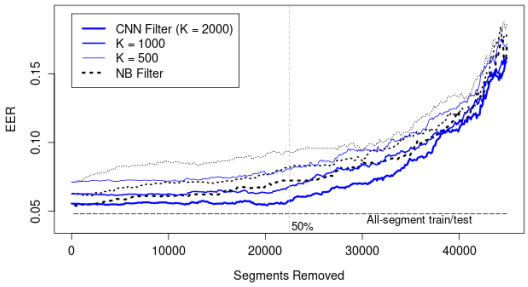

Figure 4: *Effect of KW filters on Mandarin topic detection error.*

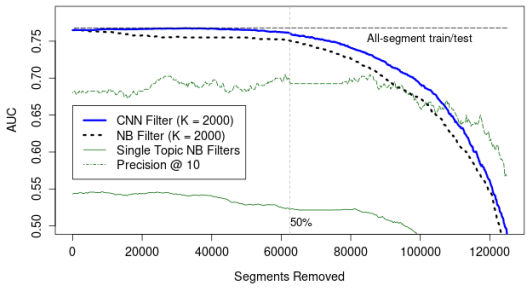

Figure 5: *Effect of KW filters on English topic precision.*

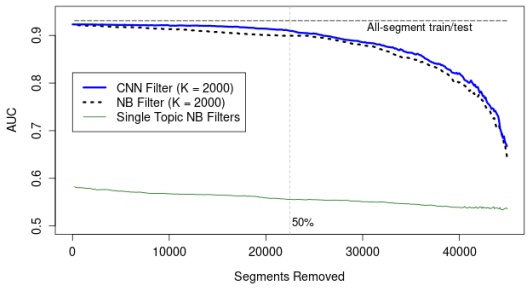

Figure 6: *Effect of KW filters on Mandarin topic precision.*

automatically from the document-level labeled training lattices.

Thus far we have treated the M-class topic ID data as a single task and reported results averaged over all M topics. Alternatively we looked at having M independent topic detectors (with M associated KW filters). Figures 5 and 6 shows the precision-based performance of this scenario as compared to the top CNN and NB multi-label based filters. We have truncated the graphs slightly to better illustrate scale across the whole range and we can see that the mAUC of the single topic filters perform significantly worse than the joint case.

For English, if we look at just the first 10 hypotheses (P@10), we do observe up to 70% precision at the top of the result list, but closer to 10% on the Mandarin task, suggesting different techniques are required for the single task scenario.

6. Conclusions

We have demonstrated that we can effectively derive topic-salient keyword detectors using only document-level labels while simultaneously learning multi-class topic classifiers. In doing so we can bring the same pre-filtering techniques for keyword search that can avoid fully decoding the majority af audio segments and reduce the computation cost of spoken topic ID without seriously degrading classification performance. Also, we demonstrate these results without relying on perfectly matched ASR acoustic models and on both English and Mandarin data sets ensuring we are not simply leveraging some English-specific phenomenon.

There is much room for further investigation, particularly in the single topic case, as well as further quantifying runtime effects, better understanding the impact of different languages or alternative topic tasks, and training our models to augment newer end-to-end frameworks.

7. References

[1] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you?" *Speech Communication*, vol. 23, no. 1-2, pp. 113–127, 1997.

[2] J. Wintrode, G. Sell, A. Jansen, M. Fox, D. Garcia-Romero, and A. McCree, "Content-based recommender systems for spoken documents," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5201–5205.

[3] M. Morchid, R. Dufour, M. Bouallegue, G. Linares, and R. D. Mori, "Theme identification in human-human conversations with features from specific speaker type hidden spaces," in *Fifteenth annual conference of the international speech communication association*, 2014.

[4] K. Church and W. Gale, "Inverse Document Frequency (IDF): A Measure of Deviations from Poisson," in *Natural language processing using very large corpora*. Springer, 1999, pp. 283–295.

[5] K. W. Church, "Empirical Estimates of Adaptation: the Chance of Two Noriegas is Closer to $p/2$ than p^2," in *Proceedings of the 18th conference on Computational Linguistics*, vol. 1. Association for Computational Linguistics, 2000, pp. 180–186.

[6] J. Wintrode and S. Khudanpur, "Limited resource term detection for effective topic identification of speech," in *Proc. of ICASSP*, 2014.

[7] J. Wintrode and J. Wilkes, "Fast Lattice-free Keyword Filtering for Accelerated Spoken Term Detection," in *Proc. of ICASSP*, 2020.

[8] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 2007, pp. 659–664.

[9] M.-H. Siu, H. Gish, A. Chan, and W. Belfield, "Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] S. Kesiraju, R. Pappagari, L. Ondel, L. Burget, N. Dehak, S. Khudanpur, J. Černockỳ, and S. V. Gangashetty, "Topic identification of spoken documents using unsupervised acoustic unit discovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5745–5749.

[11] T. J. Hazen, "Topic identification," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, vol. 12, pp. 319–356, 2011.

[12] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[13] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[14] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 427–431.

[15] C. Weng, D. L. Thomson, P. Haffner, and B.-H. F. Juang, "Latent semantic rational kernels for topic spotting on conversational speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1738–1749, 2014.

[16] C. May, F. Ferraro, A. McCree, J. Wintrode, D. Garcia-Romero, and B. Van Durme, "Topic identification and discovery on text and speech," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2377–2387.

[17] S. Kesiraju, O. Plchot, L. Burget, and S. V. Gangashetty, "Learning document embeddings along with their uncertainties," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2319–2332, 2020.

[18] V. Peddinti *et al.*, "JHU ASPiRE System: Robust LVCSR with TDNNs, ivector Adaptation and RNN-LMs," in *Proc. of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2015.

[19] Y. He *et al.*, "Streaming End-to-end Speech Recognition For Mobile Devices," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[20] P. Chitkara, A. Modi, P. Avvaru, S. Janghorbani, and M. Kapadia, "Topic spotting using hierarchical networks with self attention," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3755–3761.

[21] S. Adya *et al.*, "Hybrid Transformer/CTC Networks for Hardware Efficient Voice Triggering," in *Proc. of Interspeech*, 2020.

[22] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, 1997.

[23] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-text," in *Proceedings of LREC*, 2004, pp. 69–71.

[24] P. Fung, S. Huang, and D. Graff, "HKUST Mandarin Telephone Speech, Part 1 LDC2005S15," http://catalog.ldc.upenn.edu/LDC2005S15, Linguistic Data Consortium, 2005.

[25] X. Na, "CVTE Mandarin Model," http://kaldi-asr.org/models/m2, 2019, accessed: 2021-03-25.