MULTIPLE INDEX COMBINATION FOR JAPANESE SPOKEN TERM DETECTION WITH OPTIMUM INDEX SELECTION BASED ON OOV-REGION CLASSIFIER

Naoyuki Kanda, Katsutoshi Itoyama, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan 606-8501

naoyuki@zeus.kuis.kyoto-u.ac.jp, itoyama@kuis.kyoto-u.ac.jp, okuno@i.kyoto-u.ac.jp

ABSTRACT

In this paper, a novel index combination method for spoken term detection is proposed. In our method, outputs from four different recognizers (word, syllable, word-syllable, and fragment recognizer) are combined into one confusion network. A novel index-selection method for the multiple index-combination method is then used to suppress the increase of the index size. Two methods are proposed to reduce index size: (1) arc selection and (2) unit selection, both of which are based on an OOV-region classifier score. Experimental results with 39 hours of Japanese lecture recordings showed that the index-selection method achieved a 22% reduction of index size of the best confusion network while maintaining its high accuracy. Compared with the best phoneme-based index from a single recognizer, the proposed method achieved a 25.0% and 14.8% relative error reduction for IV and OOV queries without increasing the index size.

***Index Terms*—** Spoken term detection, keyword spotting, out-of-vocabulary detection

1. INTRODUCTION

Spoken term detection (STD) is a technique for detecting the positions where a query word or phrase is uttered in a large speech database. STD is a key module of spoken document retrieval, and many studies have been conducted, including works operated in the NIST STD [1] and NTCIR SDR [2] workshops. A simple way to realize STD is to use a large vocabulary continuous speech recognizer (LVCSR) to convert speech waves into text and use well established word-based search techniques such as an inverted index. LVCSR-based methods, however, have a defect in that they cannot detect out-of-vocabulary (OOV) queries in the LVCSR's dictionary. Many informative words such as the names of persons, names of places, and newly created words tend to be OOVs because of their scarcity; therefore, detecting OOV queries is important.

Many researchers use subword-based techniques [3, 4, 5, 6, 7] to detect OOV queries. Such techniques are often used in combination with LVCSR-based methods because their search accuracy for IV terms tends to be lower than that of LVCSR-based methods. Recently, methods that combine multiple types of indices were proposed that achieve high detection accuracy [8, 9, 10]. For example, the best performance in the latest NTCIR STD evaluation [2] was obtained by a method that combines 10 different recognizers' outputs [10]. There are many variations of index-combination methods: subword unit type (word/phoneme [4, 5], original subwords [11]), index format (lattice [4], confusion network [5, 9, 10]), and score calculation (modified edit-distance [10], weighted-sum [9]).

A defect of the multiple index-combination method is its large index size. As many indices are combined, the index size becomes larger. A larger index not only increases storage cost but also slows search speed [2, 10]. A confidence measure is often used to find redundant portions of an index made from a single recognizer; however, it is not always easy to extend this method to a combined index made from multiple recognizers because confidence measures from different recognizers are often biased differently. Furthermore, a confidence measure of the region that contains OOVs tends to have small value; therefore, confidence-measure-based index pruning may cause degradation of accuracy for OOV queries.

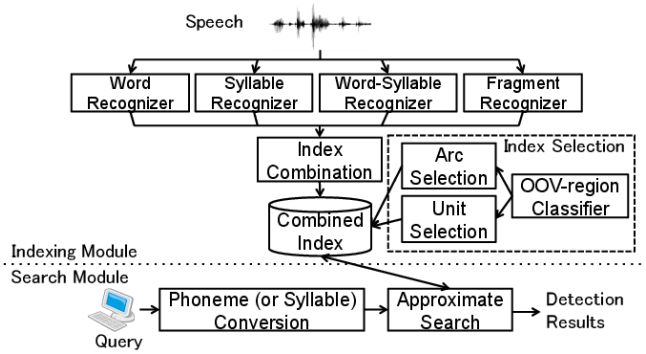


Fig. 1. Overview of spoken term detection system

In this paper, a novel index combination method for spoken term detection is proposed. In our method, outputs from four different recognizers (word, syllable, word-syllable, and fragment recognizer) are combined into one confusion network. A novel index-selection method for the multiple index-combination method is then used to suppress the increase of the index size. Experimental results with 39 hours of Japanese lecture recordings showed that the index-selection method could reduce 22% of index size of the best confusion network while maintaining its high accuracy. Compared with the best phoneme index from a single recognizer, the proposed method achieved a 25.0% and 14.8% relative error reduction for IV and OOV queries without increasing index size.

2. SPOKEN TERM DETECTION SYSTEM WITH MULTIPLE INDICES

Figure 1 depicts the system overview. The spoken term detection system consists of two modules: an indexing module and a search module. The indexing module works when new speech data is added to the system, and it makes an index optimized for spoken term detection. The search module works when a user inputs a query into the system, and it detects the positions where the keyword uttered in the speech database.

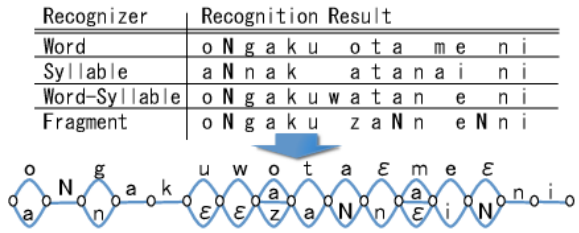


Fig. 2. Index combination as a confusion network

In the indexing module, we used four types of speech recognizers that have different language models, listed below.

Word: Word language model trained from a text corpus.

Syllable: Syllable language model trained from a corpus in which all contents are converted into syllables.

Word-Syllable: Word and syllable mixed language model trained from a corpus in which only rare words (occur less than 2 times) are converted into syllables.

Fragment: Language model trained from a fragment corpus. To make this corpus, we first prepare the syllable corpus the same as that used for the syllable language model. We then iteratively join two symbols that mostly occur successively in the corpus. The iteration is stopped when the average length of joined syllables, which we call a "fragment", becomes same as the average word length. This model can be regarded as a variation of the model proposed in [12].

Recognition results are then converted into subword sequences and combined into one index. In this paper, **syllables** or **phonemes** is used as a subword unit. We use the confusion-network combination method that Nishizaki et al. proposed [10]. This method regards the best path of a word recognizer's result as a reference sequence and then aligns other recognizers' results with the reference sequence so as to minimize the edit distance between them. Aligned sequences can be seen as a confusion network. Figure 2 depicts an example of phoneme-based index combination. Results from multiple recognizers are first aligned with each other and then combined into a confusion network. Arcs with a ε mark in the confusion network indicate epsilon transition arcs.

In the search module, we used a simple approximate search with an edit distance between the query and index. We did not incorporate a confidence measure into edit-distance calculation for simplification. Edit-distance scores are normalized by the number of subwords (phonemes or syllables) in the query.

3. INDEX SELECTION BASED ON OOV-REGION CLASSIFIER

3.1. Strategy for index selection

The aim of the index selection is to reduce index size without degrading the high accuracy of the multiple index-combination method. We introduce the OOV-region classifier to select redundant indices. The OOV-region classifier is a technique that estimates the existence probability of OOVs in an observed region [13, 14, 15]. Note that the classifier does not care which OOV-term exists[1]. We reduce an index with the two methods: the arc selection and unit selection.

The arc selection method removes index arcs originating from a specific recognizer if the OOV-region classifier score of the arcs is smaller than (or greater than) the threshold. The assumption behind this method is that some recognizers' outputs will contribute to detecting OOV (or IV) queries but will not make any contribution to IV (or OOV) query detection. In our experiments, we removed arcs if either of the following conditions, both of which were defined by preliminary experiments, were true.

1. The arc originated from the word recognizer only, and the OOV-region classifier score of the region belonged to the top $N\%$.
2. The arc originated from the syllable or word-syllable recognizer (and not from the word or fragment recognizer), and the OOV-region classifier score of the region belonged to the bottom $N\%$.

The first rule is based on an assumption that the word recognizer's output will contribute only to IV query detection. The second rule is based on an assumption that the syllable and word-syllable recognizer's outputs will contribute only to OOV query detection. We assume that the fragment recognizer will have an intermediate property and will contribute to detecting both IV and OOV queries.

The unit selection method selects an optimum subword unit for an index according to the OOV-region classifier score. The unit selection works utterance-by-utterance. The assumption behind this method is that if we know the absence of OOV in an utterance, we can use a more coarce unit for an index. In our experiment, we selected an index unit according to the rules below.

1. If the maximum OOV-region classifier score obtained from an utterance was smaller than the threshold θ, the syllable-based confusion network is used to represent the utterance.
2. Otherwise, the phoneme-based confusion network is used.

Note that most of Japanese syllables consists of two phonemes, and therefore, the syllable-based index tend to have much less arcs than the phoneme-based index. Because we need to compare detection scores from the phoneme-based and syllable-based index, we normalized the scores by using the ratio between the syllable recognition rate and phoneme recognition rate in a development data.

3.2. OOV-region classifier

We implemented an OOV-region classifier similar to that proposed by Parada et al. [14]. Bins of confusion networks provided from a word-syllable recognizer were treated as a classification unit. A conditional random field (CRF) trained with various features estimated the existence probability of OOVs in each bin of the confusion networks. We used the features below.

$$\text{Subword Existence} = \sum_{s \in t_j} p(s|t_j) \tag{1}$$

$$\text{Word Entropy} = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j) \tag{2}$$

Here, t_j indicates the current bin of the confusion networks. Variables s and w indicate syllables and words in each bin. We also used the following features: word-and-syllable-mixed entropy, best recognized word and its confidence, the difference of language-model scores from word and syllable recognizers, and the difference of acoustic-model scores from word and syllable recognizers.

4. EVALUATION

4.1. Dataset

Evaluation was conducted using 39 hours of speech from the Corpus of Spontaneous Japanese (CSJ) [16], which contains 177 recordings

[1] Many researchers call the OOV-region classifier "OOV detector" in their papers, but we think the term is confusable with detection of OOV-query in STD. Therefore, we call this technique "OOV-region classifier" in this paper.

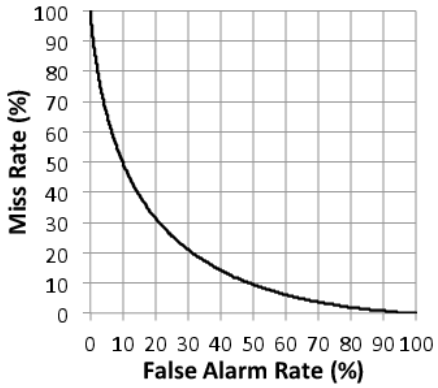


Fig. 3. Evaluation of OOV-region classifier

Table 1. Word and phoneme accuracy of speech recognizer

Recognizer	Word Acc. (%)	Phoneme Acc. (%)
Word	74.5	88.6
Syllable	-	84.9
Word-Syllable	70.8	88.7
Fragment	-	87.9

of lectures. 46 hours of speech (200 lectures) from the CSJ were used as development data to make an OOV-region classifier. The rest of the CSJ (522 hours of speech) was used as training data for an acoustic model and language models. Evaluation, development, and training data had no overlap in order to realize an open condition. A word dictionary was constructed from words that occurred more than three times in the training data. As a result, vocabulary size became 33,337, and there were 2.00% and 2.04% of OOVs in the evaluation data and development data, respectively.

A query set designed for the NTCIR-9 STD task [17], which contained 50 IV queries (occurred 14.5 times on average) and 50 OOV queries (occurred 4.7 times on average) was used. We used an F-measure (harmonic mean of precision and recall) averaged by queries as a measure of search accuracy. The detection threshold was varied and selected so as to maximize the F-measure.

Figure 3 shows the false alarm rate (FA) and miss rate (Miss) of the OOV-region classifier we implemented. The FA indicates the ratio between the number of IVs detected as OOV and the actual number of IVs. The miss indicates the ratio between the number of NOT-detected OOVs and the actual number of OOVs. For example, we could detect about 70% of the OOV-region (30% of Miss) with about 20% of false alarms.

Table 1 shows word and phoneme accuracy of the recognizers described in section 2. We used Julius [18] as the speech recognition engine. The syllable recognizer produced a slightly worse result. The other recognizers had almost the same phoneme accuracy.

4.2. Evaluation of index from single recognizer

Table 2 shows the results obtained by using the index from the single recognizer described in section 2. The column "Recognizer" shows what type of recognizer was used. "Index unit" shows what type of subword unit was used in the index. "IV" and "OOV" show the F-measure for the IV and OOV queries. "Index size" indicates the average number of arcs per one word.

Results obtained from the 1-best index and confusion network index (CN) are shown for some representative conditions. The other rows indicate results from the 1-best index. A Word CN was created

Table 2. Search accuracy and index size of single system

Recognizer	Index Unit	IV(%)	OOV(%)	Index Size
Word	Word (1-best/CN)	73.7/ **77.5**	6.7/ 9.7	1.03/ 5.90
	Syllable	75.1	35.5	1.81
	Phoneme (1-best/CN)	75.1/ 77.0	45.5/ 45.1	3.21/ 4.78
Syllable	Syllable	58.5	46.7	1.78
	Phoneme	60.8	52.8	3.16
Word-Syllable	Word	70.2	6.8	1.06
	Syllable	74.1	50.1	1.80
	Phoneme	74.9	55.1	3.20
Fragment	Syllable	67.4	43.5	1.80
	Phoneme (1-best/CN)	71.2/ 71.6	**55.4**/ 55.3	3.19/ 4.92

Table 3. Search accuracy and index size of combined system

Recognizer	Index Unit	IV(%)	OOV(%)	Index Size
All	Syllable	**81.6**	56.1	2.35
	Phoneme	80.6	**62.6**	3.87
All w/o W	Phoneme	79.3	**63.2**	3.84
All w/o S	Phoneme	80.2	60.2	3.63
All w/o WS	Phoneme	80.3	60.4	3.84
All w/o F	Phoneme	78.2	60.3	3.73

W: Word, S: Syllable, WS: Word-Syllable, F: Fragment

in the same manner described in the paper [5]. A Phoneme CN was created from the 5-best hypotheses by using the method described in section 2. As shown in table 2, the confusion networks did not always improve accuracy. This would be because the approximate matching that we used produced too many false positives by using the confusion network.

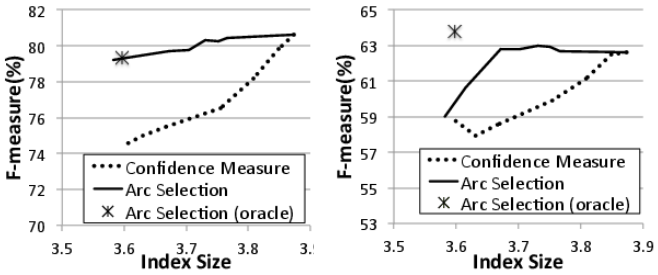
The best F-measure for the IV queries was obtained by using the word confusion network (CN) created from the word recognizer's output (77.5%). However, as expected, this method got a poor F-measure (9.7%) for the OOV queries[2]. The best F-measure for the OOV queries was obtained by using the phoneme-based index from the fragment recognizer (55.4%). The phoneme-based index from the word recognizer produced much worse results (45.5%). Compared with the syllable-based index, the phoneme-based index always produced more accurate results. Instead, it had about 3.2 arcs per a word, which was always larger than the syllable-based index.

4.3. Evaluation of index combination

Table 3 shows the results obtained by using the combined confusion network. The first column shows which recognizer's outputs were combined. The remaining columns are the same as those in table 2.

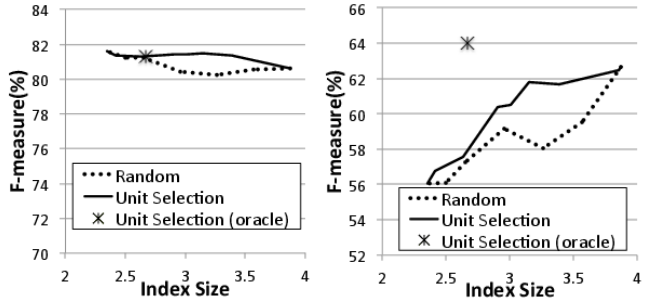
The syllable confusion network showed improvement, especially for IV queries (81.6% of F-measure); however, accuracy for OOV queries was still low. The phoneme confusion network achieved high F-measures for both IV and OOV queries (80.6% and 62.6%, respectively); however, it had a relatively large index size (3.87 arcs per a word). Interestingly, the best F-measure for the OOV queries (63.2%) was obtained from the phoneme confusion network without using a word recognizer. This result suggested that

[2]Some compound words in OOV queries are designed to consist of OOV and IV words, and they were detected by using an approximate search. This is why the word-based method got an F-measure greater than 0.



(a) IV query (b) OOV query

Fig. 4. Evaluation of arc selection method



(a) IV query (b) OOV query

Fig. 5. Evaluation of unit selection method

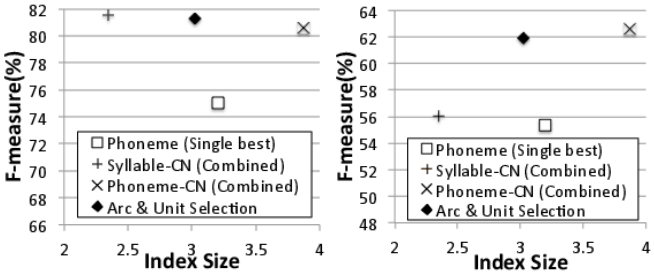
the index from the word recognizer only increased false positives when detecting OOV queries. With a similar analysis, the syllable recognizer and word-syllable recognizer seemed to contribute little to detecting IV queries. The fragment recognizer was promising: it improved accuracy for both IV and OOV queries. Note that the same trends were observed in experiments with development data, but we omitted the results due to a page limitation.

4.4. Evaluation of index selection

In this evaluation, we focused especially on the phoneme confusion network and reduced its size according to the index-selection method described in section 3. We first evaluated the arc selection method. F-measures with various points of N are shown in figure 4. The confidence-measure-based method, which removed arcs whose confidence measure[3] was smaller than the threshold, was evaluated as a reference (shown as "Confidence Measure"). We also evaluated the arc selection based on the oracle OOV-region classifier that provided totally correct results (shown as "oracle"). We first observed that the confidence-measure-based method severely degraded accuracy even for IV terms as arcs were removed. This result suggested the difficulty of using confidence measures from different recognizers. The proposed arc selection method worked better and could remove 3.7% of arcs without degrading accuracy. The oracle OOV-region classifier provides much better results, especially for OOV queries.

Next, the unit selection method was evaluated. F-measures with various points of threshold θ are shown in figure 5. Because we cannot find previous works that obviously relate to the unit selection method, we evaluated random selection of the index unit (shown as "random") as reference. Unit selection based on the oracle OOV-region classifier (shown as "oracle") were also evaluated. The OOV-region classifier-based method showed much better results than random selection, and it achieved an 18.7% reduction of index size with slight degradation of accuracy. The oracle OOV-region classifier again produced a much better result for OOV queries.

[3]If arcs originated from multiple recognizers, the maximum confidence measure was used.



(a) IV query (b) OOV query

Fig. 6. Evaluation of mixed method

Finally, we combined the arc selection method (at index size was 3.74) and unit selection method (at index size was 3.15). Figure 6 shows the results from various systems. The proposed method got a 81.3% and 61.9% F-measure for IV and OOV queries, respectively. At this point, the proposed method had 3.02 arcs per one word, which corresponded to a 22% reduction of the index size from the combined phoneme confusion network. Compared with the phoneme-based index from a single recognizer (maximum 75.1% for IV and 55.4% for OOV at the point of similar index size), the proposed method improved the F-measure by 6.2 and 6.5 points (25.0% and 14.8% relative error reduction, respectively) without increasing index size.

5. RELATION TO PRIOR WORK

The work presented here is focused on the muliple index combination method. Our index-combination method is based on Nishizaki's method [10] that combines 10 different recognizers' outputs. In this paper, we newly introduced the fragment language model into mulitple index combination method and the fragment model showed very promising results in the experiments. We also introduced a novel index selection method to suppress the increase of the index size.

There were many studies that proposed combining a word index and subword index [3, 4, 5]. For such indices, it was obvious that the subword-index was redundant for the region where words were correctly recognized, and there were some studies on pruning subword index according to the IV-word existence score (ex. word posteriori probability) [5, 6]. Proposed index-selection method can be regarded as the extention of above works to the state-of-the art multiple index combination method [8, 9, 10], for which an obvious index-selection method did not exist.

6. CONCLUSION

In this paper, a novel index combination method for STD was proposed. Outputs from four different recognizers (word, syllable, word-syllable, and fragment recognizer) were combined. Two index-selection methods based on OOV-region classifier were then introduced, and they achieved a 22% reduction in index size while maintaining the high accuracy of the combined-index. Compared with the best phoneme-based index from a single recognizer, the proposed method achieved a 25.0% and 14.8% relative error reduction for IV and OOV queries without increasing the index size.

7. REFERENCES

[1] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51–55.

[2] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for spoken documents task in NTCIR-9 workshop," in *Proc. NTCIR-9*, 2011.

[3] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. HLT-NAACL*, 2004, pp. 129–136.

[4] P. Yu and F. Seide, "A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech," *Proc. ICLSP04*, 2004.

[5] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–73.

[6] N. Kanda, H. Sagawa, T. Sumiyoshi, and Y. Obuchi, "Open-vocabulary keyword detection from super-large scale speech database," in *Proc. MMSP*. IEEE, 2008, pp. 939–944.

[7] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," in *Proc. ICASSP*. IEEE, 2010, pp. 5286–5289.

[8] S. Meng, P. Yu, J. Liu, and F. Seide, "Fusing multiple systems into a compact lattice index for chinese spoken term detection," in *Proc. ICASSP*. IEEE, 2008, pp. 4345–4348.

[9] I. Bufyko, O. Kimball, M.H. Siu, J. Herrero, and D. Blum, "Detection of unseen words in conversational mandarin," in *Proc ICASSP*. IEEE, 2012, pp. 5181–5184.

[10] H. Nishizaki, H. Furuya, S. Natori, and Y. Sekiguchi, "Spoken term detection using multiple speech recognizers outputs at NTCIR-9 SpokenDoc STD subtask," in *Proc. NTCIR-9*, 2011.

[11] Y. Itoh, K. Iwata, M. Ishigame, K. Tanaka, and S. Lee, "Spoken term detection results using plural subword models by estimating detection performance for each query," in *Proc. INTERSPEECH*, 2011.

[12] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Proc. INTERSPEECH*, 2005, pp. 53–56.

[13] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," in *Proc. ICASSP*. IEEE, 2009, pp. 3953–3956.

[14] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *Proc. NAACL-HLT*. Association for Computational Linguistics, 2010, pp. 216–224.

[15] L. Qin, M. Sun, and A. Rudnicky, "System combination for out-of-vocabulary word detection," in *Proc. ICASSP*. IEEE, 2012, pp. 4817–4820.

[16] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[17] Y. Itoh, H. Nishizaki, X. Hu, H. Nanjo, T. Akiba, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita, and K. Aikawa, "Constructing japanese test collections for spoken term detection," in *Proc INTERSPEECH*, 2010.

[18] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.