PROCEEDINGS

ICASSP 91

1991 INTERNATIONAL CONFERENCE ON
ACOUSTICS, SPEECH, AND SIGNAL PROCESSING
MAY 14-17, 1991

IEEE SIGNAL
PROCESSING SOCIETY

IEEE
91CH2977-7

VOLUME 1:

S_1 SPEECH PROCESSING 1

Lennig

ICASSP 91

VOLUME 1

S_1

SPEECH PROCESSING 1

1991
INTERNATIONAL CONFERENCE
ON
ACOUSTICS, SPEECH, AND
SIGNAL PROCESSING

MAY 14-17, 1991
THE SHERATON CENTRE HOTEL & TOWERS
TORONTO, ONTARIO, CANADA

IEEE
91CH2977-7

Sponsored by
THE INSTITUTE OF ELECTRICAL
AND ELECTRONICS ENGINEERS,
SIGNAL PROCESSING SOCIETY

Using phoneme duration and energy contour information to improve Large Vocabulary Isolated-word Recognition[1]

V.N. Gupta, M. Lennig*, P. Mermelstein*, P. Kenny, F. Seitz, and D. O'Shaughnessy*
INRS–Télécommunications, Montreal, Quebec, Canada

Abstract

Many acoustic misrecognitions in our 86,000-word speaker-trained isolated word recognizer are due to phonemic hidden Markov models (phoneme models) mapping to short segments of speech. When we force these models to map to larger segments corresponding to the observed minimum durations for the phonemes, then the likelihood of the incorrect phoneme sequences drops dramatically. This drop in the likelihood of the incorrect words results in significant reduction in the acoustic recognition[2] error rate.

We have also improved the phoneme models by correcting the segmentation of the phonemes in the training set. During training, the boundaries between phonemes are not marked accurately. We use energy to correct these boundaries. Application of an energy threshold improves the segment boundaries between stops and sonorants (vowels, liquids, and glides), between fricatives and sonorants, between affricates and sonorants, and between breath noise and sonorants. Training the phoneme models with these segmented phonemes results in models which increase recognition accuracy significantly. On two speakers, the overall reduction in errors using minimum durations and energy thresholds is from 27.3% to 23.1% for acoustic recognition, and from 14.3% to 8.8% with the language model.

I. Introduction

The goal of our 86,000-word recognizer is to transcribe speech spoken as a sequence of isolated words. For each spoken word, the recognizer uses acoustic information and rough likelihoods in a fast search algorithm [Gupta, Lennig and Mermelstein, 1988] to narrow the possible word hypotheses from the 86,000 words [Seitz et al., 1990] in the total vocabulary to a small list. It then refines the list by computing an exact likelihood score for each hypothesized word. The exact likelihood scores take into account acoustic information but not the syntactic, semantic, and pragmatic characteristics of English. To take these into account, the exact likelihoods are further refined with the aid of a statistical language model to generate the most likely sequence of words [Gupta, Lennig and Mermelstein, 1991]. In this paper, our focus is on improving the accuracy of our recognizer through improvements in the acoustic recognition algorithm.

Our strategy is to improve recognition accuracy by eliminating weaknesses inherent in hidden Markov modeling algorithms for speech recognition. For instance, hidden Markov models incorporate only weak duration constraints in the phonemes they generate. During both the fast search algorithm and the exact likelihood scoring, the phoneme[3] models are mapped to acoustic segments in order to compute the likelihood of the acoustic data. When an incorrect phoneme sequence is mapped to the acoustic input, we frequently observe that one or more models are often mapped to acoustic segments shorter than minimum possible duration for the phoneme. One such example can be seen in Fig. 1.

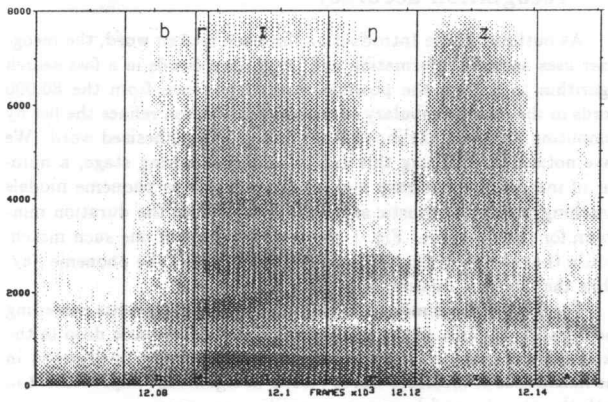

Fig. 1 An example of an HMM mapping to a short duration acoustic segment

Here, the word spoken is *veins*, while the best choice produced by the recognizer is *brings*. Notice that the model for /r/ is mapped to 20 ms of acoustic data, while the *duration minimum*[4] for /r/ observed in the training data is 40 ms. We consider mapping of phoneme models to acoustic segments shorter than those observed in the training data as a weakness in hidden Markov modeling. This weakness has been addressed by imposing minimum duration constraints on the phonemes in the HMM framework.

Another weakness in hidden Markov modeling that we have addressed is the erroneous segmentation of words into phonemes. Many segmentation errors in the training data are between high energy and low energy phonemes. A number of these segmentation problems have been corrected by constraining the energy contours to prevent phonemes with high energy from mapping onto phonemes with lower energy. Segment boundaries between stops and sonorants, between fricatives and sonorants, between affricates and sonorants, and between breath and sonorants are most amenable to correction using energy constraints. Segment boundaries between vowels, liquids, glides, and nasals can be corrected by using duration minima. Correction of the segment boundaries between phonemes in the training set leads to improved phoneme models, resulting in higher acoustic recognition accuracy and higher recognition accuracy with the language model.

In Section II, we give details on how we exploit the phoneme duration minima to improve the accuracy of our recognizer. Improvements in phoneme segmentation in the training set through energy thresholds is discussed in Section III.

[1] This work was supported by the Natural Sciences and Engineering Research Council of Canada.

* Also with Bell-Northern Research, Montreal.

[2] We use the term acoustic recognition error rate to mean the recognition error rate when every word in the vocabulary is considered *a priori* equally likely.

[3] The models represent phonemes, except for /l/ and /r/, where we use two allophones (prevocalic and syllabic/postvocalic). Despite the fact that /l/ and /r/ each have two models, we refer to each of the 44 models in our system as a phoneme model.

CH2977-7/91/0000-0341 $1.00 © 1991 IEEE

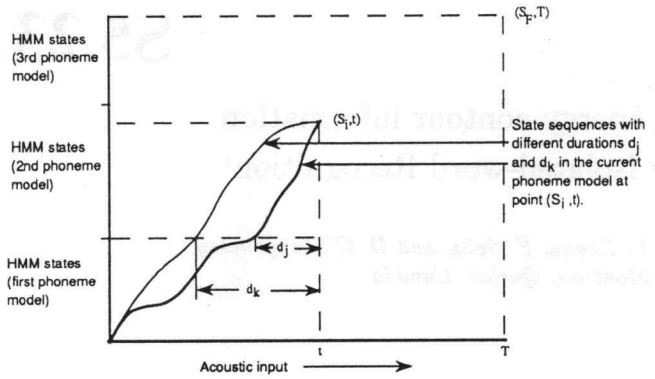

Fig. 2 Trellis for Viterbi search showing how duration is incorporated in the search.

II. Exploiting phoneme durations to improve recognition accuracy

As outlined in the Introduction, for each spoken word, the recognizer uses acoustic information and rough likelihoods in a fast search algorithm to narrow the possible word hypotheses from the 86,000 words in the total vocabulary to a small list. It then refines the list by computing an exact likelihood score for each hypothesized word. We have noticed that during this exact likelihood scoring stage, a number of incorrect words have high likelihoods due to phoneme models matching well with acoustic segments shorter than the duration minimum for the phoneme. Fig. 1 shows an example of one such match. Notice that the model for [r] is mapped to 20 ms of the phoneme /e/, while the duration minimum for [r] is 40 ms.

Forcing phoneme models to match to acoustic segments exceeding the duration minima of the phonemes results in dramatic drop in the likelihoods of many of these incorrect word choices. Such drops in the likelihoods of incorrect words results in significant improvement in both the acoustic and language recognition accuracy.

There are many possible strategies for incorporating duration constraints[5]. One possibility is to estimate separately the probability distribution of the phoneme durations from the training data and use them during recognition [Lee and Rabiner, 1989] or they can be applied in a postprocessor [Rabiner, Wilpon and Juang, 1987]. A second possibility is to structure the phoneme models so that the minimum number of transitions necessary to traverse the model corresponds to the duration minimum for the phoneme (this will not allow context dependent duration constraints). A third possibility is to constrain the possible state sequences through the model to correspond to the duration minimum for the phoneme. We have chosen the third alternative. Our goal here is to improve recognition accuracy by addressing weaknesses in hidden Markov modeling, and not to compare different ways of modeling durations. These duration constraints have been applied to improve our recognition system which uses phoneme models[6] [Deng et al., 1990].

Consider how these duration constraints can be incorporated into the HMM framework. We would like to find the most likely state sequence (through the phoneme models corresponding to the given phoneme sequence) which generates the given acoustic input and obeys the duration minima for the phonemes. To obtain such a state sequence, let us look at all possible paths which end at state S_i at time t (see Fig. 2).

In the Viterbi formulation without duration constraints, we only keep the most likely state sequence to state S_i at time t. To incorporate duration constraints, we also compute the duration in frames of the current phoneme up to state S_i at time t. This duration d_j corresponds to the total number of transitions taken through the current model in order to reach state S_i at time t. To impose minimum duration constraints, we do not allow a transition from state S_i at time t to a state in the model for the next phoneme if the duration d_j is less than the minimum allowed duration.

If we keep only the best state sequence for each state S_i at time t, then the resulting path (or state sequence) to the final state S_F at time T will be suboptimal. To see this, let us assume that d_j is less than the duration minimum. Therefore, this path cannot be extended to the next model (by taking a transition from state S_i to a state in the following model). Also, any extension of this path to any state S_j in the current phoneme model may not correspond to the most likely path to state S_j. In other words, this path will be lost. However, if there was another path to state S_i at time t with a lower likelihood which could be extended to the next phoneme (i.e., the path satisfies minimum duration constraints), then this path may eventually become the partial path leading to the most likely state sequence to the final state S_F at time T. Therefore, to determine the optimal path, multiple paths must be maintained at each state S_i at time t.

Let us see how many paths need to be kept at each point determined by state S_i and time t. Note that the multiple paths correspond to the most likely state sequences with different durations in the current model. If a path with duration greater than the minimum has the highest likelihood, then all other paths at that point can be discarded (we only need to keep one path in such cases). Also, at every point, we need to keep only one path with duration greater than the minimum. If a path with duration greater than the minimum exists, then we need only keep paths with duration less than the minimum which have likelihoods greater than this path.

In deriving the duration minimum, we have looked at the durations of phonemes in the training set for two speakers with fast speaking rate (among the nine speakers). We have derived one duration minimum per phoneme, except for stops and fricatives, whose duration minimum depends on position. For each stop and fricative, we use three possible duration minima corresponding to the phoneme occurring in initial, medial or final position[7]. Stops in initial position can be shorter, since the silent stop gap or voice bar may not be present. Similarly, final stops may be unreleased or very weakly released, in which case the endpoint detector may not classify the stop burst as part of the word. Also, initial and final weak fricatives /θðfv/ can be shorter. Therefore, we have reduced the minimum durations for initial and final weak fricatives as shown in Table 1. These phoneme durations observed are generally longer than what would be expected in continuous speech. The minimum durations thus derived are used for all speakers.

Recognition results with and without duration constraints are compared for nine speakers in Table 2. The duration constraints reduce the number of search errors in the fast search algorithm. A search error occurs when the correct word is absent from the list of possible word hypotheses. The number of search errors are reduced from 210 to 121 (3.1% to 1.8%). Application of duration constraints to the exact likelihood scoring stage results in a reduction of acoustic recognition errors from 18.6% to 17.3%. However, the major benefit of the duration constraints becomes evident after applying the language model. There, the duration constraints reduce the word errors by 22% (from 9.2% to 7.2%). The reason for the reduced error rate is the relative lowering of the likelihoods of incorrect words as compared to the correct word. This is evident from Table 3 which shows the average values of (log likelihood (correct word) – log likelihood (incorrect word)) with and without duration constraints.

[4] We refer to the minimum duration of a phoneme observed in the training set as its duration minimum.

[5] The models already contain weak duration constraints through state transition probabilities. These duration constraints do not penalize short durations adequately.

[6] These phoneme models are left-to-right Gaussian mixture HMMs with self loop, next state, and skip transitions.

[7] We use the terms initial, medial, and final to mean word-initial, word-medial, and word-final, respectively.

allophone	dur	allophone	dur	allophone	dur
/aj/	100 ms	/aw/	100 ms	/ɔj/	100 ms
/i/	60 ms	/ɪ/	40 ms	/ɑ/~/ɔ/	70 ms
/ɛ/	70 ms	/ɛ/	40 ms	/æ/	60 ms
/ɑr/	100 ms	/ʌ/	60 ms	/ʊ/	50 ms
/u/	70 ms	/o/	70 ms	/ɔr/	80 ms
/ə/	20 ms	/j/	30 ms	/w/	50 ms
[l]	40 ms	[l̩]	60 ms	[r]	40 ms
[r̩]	60 ms	initial breath	30 ms	final breath	30 ms
/p/	60 ms	/t/	40 ms	/k/	70 ms
/b/	40 ms	/d/	40 ms	/g/	40 ms
/tʃ/	80 ms	/dʒ/	70 ms	/f/	70 ms
/v/	40 ms	/θ/	70 ms	/ð/	50 ms
/s/	90 ms	/z/	80 ms	/ʃ/	100 ms
/ʒ/	60 ms	/m/	40 ms	/n/	40 ms
/ŋ/	70 ms	/h/	60 ms		
initial [t b d g ð]					20 ms
initial [p f θ v]					40 ms
initial [k]					60 ms
final [p b t d k g f θ v ð]					20 ms

Table 1 Duration minima in milliseconds for different phonemes used in recognition. Phonemes in the last four rows have position dependent minima.

spkr	total words		acoustic recognition				errors after lang model	
			search err.		recog err.			
(sex)	test	train	no dur	dur	no dur	dur	no dur	dur
DS(m)	451	1900	4.9	3.1	24.0	21.9	14.4	10.4
AM(m)	565	2742	3.6	1.8	30.6	31.0	14.2	12.2
ML(m)	596	2000	1.7	1.2	14.5	12.6	6.7	5.4
JM(m)	587	1664	3.0	3.0	23.9	22.7	8.2	7.8
FS(m)	1014	2322	1.7	1.3	8.4	7.5	5.0	3.7
NM(f)	967	1299	4.0	1.7	19.4	15.0	11.0	6.0
CM(f)	1090	2343	3.5	2.1	16.9	16.5	8.9	7.8
MM(f)	586	2338	2.2	1.4	14.3	14.3	5.0	3.6
LM(f)	863	2353	3.8	1.7	23.7	22.6	12.1	9.8
Avg	6719	2107	3.1	1.8	18.6	17.3	9.2	7.2

Table 2 Recognition results (in percentage) for nine speakers with and without duration constraints.

speaker	log like(correct) – log like (incorrect)	
	no constraints	duration constraints
LM	14.6	18.0
ML	20.1	29.8
NM	20.5	31.2

Table 3 Comparison of (log likelihood(correct word) – log likelihood (top choice incorrect word)) before and after duration constraints

III. Use of energy measure to improve training of phoneme models

The phoneme models can be trained by forward-backward algorithm or by Viterbi algorithm [Levinson et al., 1983]. In the context of our large vocabulary recognizer, the two training algorithms result in identical recognition accuracy. With Viterbi training, the implied segmentations can be evaluated to see how effectively the phoneme models segment words into phonemes. We have observed many phoneme segmentation errors in the training data, and we consider this a weakness of hidden Markov modeling. By correcting the segment boundaries between phonemes in the training set, we can improve the phoneme models and enhance the accuracy of the recognizer. Both duration and energy constraints are helpful in correcting such segmentation errors.

In the training data, many segmentation errors are between low energy and high energy phonemes. For example, we have observed incorrect segment boundaries between stops and sonorants, between fricatives and sonorants, between affricates and sonorants, and between breath noise and sonorants. (Note that stops, fricatives, affricates and breath have low energy, while sonorants have high energy). Segment boundaries between low energy and high energy phonemes can be located accurately by forcing a more precise alignment of energy contours. One example of incorrect segment boundaries can be seen in the top spectrogram of Fig. 3 which shows Viterbi segmentation without any duration or energy constraints.

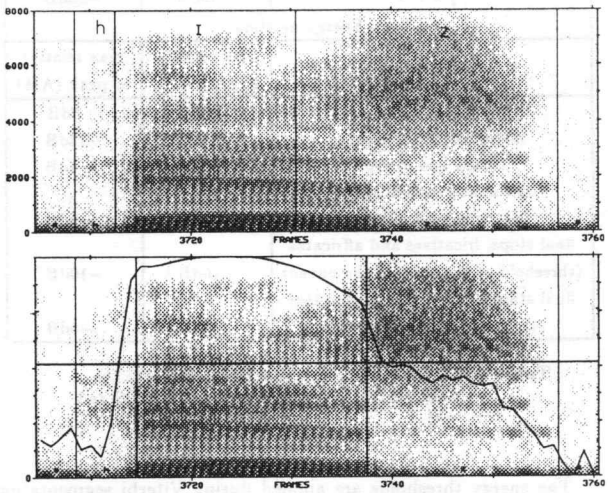

Fig. 3 Example of improved segmentation using duration and energy constraints in training.

In this example, the boundaries between /h/ and /ɪ/, and between /ɪ/ and /z/ in *his* are not right. In this instance, we can use the frame energy (C_0) to correct the segmentation errors. The segment boundaries after consideration of energy thresholds are shown in the bottom spectrogram of Fig. 3. The energy contour is superimposed on the spectrogram. The energy increases rapidly from /h/ to /ɪ/, and it drops rapidly during transition from /ɪ/ to /z/.

Sonorants have similar energy contours, therefore energy thresholds cannot be used to correct segment boundaries between them. However, minimum duration constraints can be used to correct segment boundaries between sonorants. In other words, the energy and duration constraints are complementary. The duration constraints are imposed in the same way as in recognition as outlined in Section II.

We have observed that the C_0 thresholds required to achieve the best segment boundaries are speaker dependent. The thresholds were optimized iteratively from the training data by using a set of initial threshold assignments, looking at the resulting segmentation, and then selecting new threshold values to improve the segmentation. The thresholds used for speakers DS and AM are compared in Table 4.

Let us discuss some of the thresholds given in Table 4 in more detail. In all vowels (except /ɪ/ and /ə/), the lowest energy in any frame inside a vowel can drop as low as 15 dB below the *peak energy*, where the peak energy in the word is that of the frame with the highest energy. The lowest energy in /ɪ/ and /ə/ can be as low as 20 dB below the peak energy for speaker DS, and as low as 25 dB below the peak

energy for speaker AM. Note that, as the threshold constraints get weaker, the effectiveness of thresholds to correct segmentation errors is reduced. In fact, the energy threshold is only marginally effective in correcting segmentation errors between the final vowel and a non-sonorant phoneme for speaker AM. This is primarily due to the strong vocal fry observed in the final syllable for this speaker. Also, as is evident from Table 4, the energy constraints are only marginally effective in placing a segment boundary between a flap and a sonorant segment.

Energy minima		
phonemes	min relative to peak (DS)	min relative to peak (AM)
/aj aw ɔj ɑ ɔ i e ɛ æ ɑr ʌ u ʊ o ɔr/	−15dB	−15dB
/ɪ ə/	−20dB	−25dB
final vowel	−15dB	−35dB
/j w l r/	−30dB	−35dB
Energy maxima		
phonemes	max relative to peak (DS)	max relative to peak (AM)
/n m ŋ/	0dB	0dB
initial and final breath	−15dB	−15dB
t-flaps and d-flaps	−1dB	−1dB
initial and medial stops, fricatives, and affricates	−5dB	−5dB
final stops, fricatives and affricates (threshold at the start of the phone)	−10dB	−10dB
final stops, fricatives and affricates (threshold everywhere else)	−5dB	−5dB

Table 4 C_0 constraints for phonemes (relative to the peak energy in the word) for speakers DS and AM.

The energy thresholds are applied during Viterbi segmentation of words into phonemes in the training set. Application of energy thresholds is very similar to the application of duration constraints as explained in Section II. In applying energy constraints, a transition in the phoneme model can only be taken if the C_0 value for the observed frame is within the specified constraints for the corresponding phoneme model. Such a restriction gives us Viterbi segmentation of the word into phonemes with the specified constraints on C_0 (see Table 4 for C_0 constraints). The new phoneme models are trained using this segmented training data.

Recognition results for the two speakers using the new phoneme models are shown in Table 5. The recognition accuracy for both the speakers has improved significantly. The acoustic recognition errors have been reduced by 13%, while the recognition errors after the language model have been reduced by 23%. For the speakers DS and AM, the combined effect of duration constraints in recognition, and duration and energy constraints in training is a 16% reduction in acoustic recognition errors and a 39% reduction in errors after the language model.

IV. Conclusions

We have used minimum duration constraints and energy thresholds for phonemes to increase the recognition accuracy of our large vocabulary recognizer. Minimum duration constraints force the phoneme models to map to acoustic segments longer than the duration minima for the phonemes. Such minimum duration constraints result in significant lowering of likelihoods of many incorrect word choices, improving the accuracy of both acoustic recognition and recognition with the language model. The number of word recognition errors with the language model is reduced from 620 to 481 (9.2% to 7.2%). The minimum

spkr	acoustic recog				errors after lang model	
	search errors		recog errors			
	no C_0	C_0	no C_0	C_0	no C_0	C_0
DS	3.1%	3.0%	21.9%	18.8%	10.4%	7.5%
AM	1.8%	2.2%	31.0%	27.3%	12.2%	10.0%
Avg	2.5%	2.6%	26.5%	23.1%	11.3%	8.8%

Table 5 Recognition results with and without energy (C_0) and duration constraints during training. During recognition, duration constraints are used.

phoneme duration constraints applied are same across all nine speakers tested, so the durations do not have to be trained to individual speakers.

Energy thresholds improve segment boundaries between phonemes in the training set. Training the phoneme models with improved segmentation for phonemes results in increased acoustic recognition accuracy and in enhanced recognition accuracy after the language model. Not all segment boundaries between phonemes can be improved by using energy thresholds. Only the segment boundaries between stops and vocalic segments, between fricatives and vocalic segments, between affricates and vocalic segments, and between breath and vocalic segments can be improved. Also, the energy thresholds used are speaker dependent and cannot be applied blindly across all speakers. Vocal fry significantly reduces the effectiveness of the energy thresholds. The energy thresholds are tighter for the speaker with very little vocal fry than for speaker exhibiting a lot of vocal fry. The combined effect of minimum duration constraints in recognition and minimum duration and energy constraints in training is to reduce the word error rate by 40% (from 14.3% to 8.8%) with the language model. To achieve speaker-independent thresholds, one would have to examine additional data from numerous speakers. Presumably, a uniform composite set of thresholds could yield results not much worse than the set optimized individually for each speaker.

References

Deng, L., Lennig, M., Gupta, V., Kenny, P., Seitz, F. P. and Mermelstein, P. (1990). "Acoustic Recognition Component of an 86,000-word Speech Recognizer," *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing*, 741–744.

Gupta, V., Lennig, M., and Mermelstein, P. (1988). "Fast search strategy in a large vocabulary word recognizer," J. Acoust. Soc. Am., **84**(6), 2007–2017.

Gupta, V., Lennig, M., and Mermelstein, P. (1991). "A language model for very large vocabulary speech recognition," submitted to IEEE Transactions on Acoustics, Speech, and Signal Processing.

Lee, C.-H. and Rabiner, L.R. (1989). "A frame-synchronous network search algorithm for connected word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, **ASSP-37**(11), 1649–1658.

Levinson, S.E., Rabiner, L.R. and Sondhi, M.M. (1983). "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," Bell Systems Technical Journal, **62**(4), 1035–1074.

Rabiner, L.R., Wilpon, J.G., and Juang, B.-H. (1987). "A performance evaluation of a connected digit recognizer," *Proceedings of the 1987 IEEE International Conference on Acoustics, Speech and Signal Processing*, 101–104.

Seitz, F., Gupta, V., Lennig, M., Kenny, P., Deng, L., and Mermelstein, P. (1990). "A dictionary for a very large vocabulary speech recognition system," Computer, Speech and Language, **4**, 193–202.

Soong, F.K. (1989). "A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis," *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing*, 584–587.