MODELING LONG-RANGE DEPENDENCIES IN SPEECH DATA FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Ji Ming[a], *Jie Lin*[a,b]

[a]Institute of ECIT, Queen's University Belfast, Belfast BT7 1NN, UK
[b]School of Computer Science, University of Electronic Science and Technology, Chengdu, China

ABSTRACT

In the paper, a new approach for modeling and matching long-range dependencies in free-text speech data is proposed for speaker recognition. The new approach consists of a sentence model to detail up to sentence-level dependencies in the training data, and a search algorithm that is capable of locating the matches of arbitrary-length segments between the training and testing sentences. The search algorithm is optimized to increase the probability for the match of long, continuous segments as opposed to short, separated segments, assuming that long, continuous segments contain more specific information about the speaker. The new approach has been evaluated on the NIST 1998 Speaker Recognition Evaluation database, and has shown improved performance.

Index Terms— Time dependence, segment modeling, speaker modeling, speaker recognition

1. INTRODUCTION

Gaussian mixture models (GMMs) have been a major approach to speaker recognition. Built on statistics, a GMM has the ability to smoothly, and therefore robustly, represent the probability distribution of a speaker's sounds, usually at a time range of 10-20 ms (i.e., a frame). The GMMs, however, lack the ability to show how these short-time sounds are dependent on one another to form a real-world speech sentence. In recent years, many approaches have been studied for modeling the dependencies between short-time acoustic features for speaker recognition. It is assumed that both the short-time features and their dependencies, especially over long time ranges, carry information about the speaker.

Many studies have considered to label the speech signals into longer acoustic segments, such as broad phonetic classes or quasi-stationary segments. Each class or segment is then modeled and classified individually as a single unit. These approaches try to capture the correlation within each class/segment, which typically consists of several frames. HMMs (e.g., [1], [2]), neural networks [3], and ALISP segmentation [4] have been used as the labeller and classifier. A similar approach is described in [5], in which a large-vocabulary continuous speech recognition (LVCSR) system is used to segment the input speech into pre-defined acoustic units, and this is followed by GMMs one for each unit group. Other studies have considered the incorporation of duration or prosodic features, such as pitch patterns, energy trajectories, etc., into the recognition process (e.g., [6]–[8]).

More recent studies have re-considered the use of templates to capture dependencies in the speech signals. For example, an approach is described in [9] in which a LVCSR system is used to identify similar phonetic units between the training and testing data, both represented in templates; the acoustic similarity of the identified units is subsequently decided by using a DTW (dynamic time warping) algorithm, which produces the score for recognition. A similar approach is presented in [10], which directly uses DTW to spot and compare similar words between the training and testing sentences. In contrast to other modeling approaches, templates involve less manipulation on the speech data and, thus, may be able to more accurately represent the dependencies in the given signals. However, unlike other statistical models, templates lack smoothness (and hence robustness) in representing the short-time features, which are subjected to random variations.

In this paper, we propose a new approach to address the problem. The new approach has the following characteristics: (1) it combines statistical and example-based approaches seamlessly in the same framework, to offer both smooth representation for the short-time features and sentence-long representation for the dependencies; (2) it allows the match of arbitrary-length segments between the training and testing data, not limited to any subword or word units, and not limited to linguistical segments; and (3) it combines the segmentation and recognition in the same framework subject to a joint optimality criterion – to focus the recognition on the longest matching segments. Long segments usually contain richer dependencies and thus, should be emphasized for likely carrying more specific information about the speaker.

2. METHODOLOGY

2.1. Modeling Long-Range Dependencies

The new approach consists of two parts: (1) a seed acoustic model representing the short-time features of all the training sentences, and (2) a set of sentence models, built on the seed model and one sentence model for each training sentence, representing the dependencies of short-time features in the training data. The sentence models of the same speaker are grouped together and used for recognition. We start to build the new system by training a seed model with the available training data. The seed model could be either a traditional unsupervised representation such as a GMM, or a traditional supervised representation such as the phone-class models (each model itself is usually a GMM or HMM). Many of the previous approaches to speaker modeling have stopped at the construction of the seed model, which characterizes a speaker's sounds at the time range of a frame or a phone. In this research, we move one step further. We add models for long-range dependencies, up to the length of sentences, for recognition.

Without loss of generality, we assume that a GMM is used as the seed model for each speaker. Denote the GMM for speaker λ as

$$G_\lambda = \{p_\lambda(x|k), w_\lambda(k) : k = 1, 2, ..., K\} \quad (1)$$

where $p_\lambda(x|k)$ represents the k'th Gaussian component and $w_\lambda(k)$ is the corresponding weight. G_λ models the probability distribu-

tion of a speaker's short-time features, as seen in the training data. Based on G_λ, we further develop a representation for each training sentence that captures the full-time dependencies within the sentence. Let $X = (x_1, x_2, ..., x_T)$ be a training sentence for speaker λ, with T frames and x_t being the frame at time t. A probabilistic representation for X can be obtained by taking each frame from X and finding the Gaussian component in G_λ that maximizes the likelihood of the frame. This results in a time sequence of likelihoods $\{p_\lambda(x_1|k_1), p_\lambda(x_2|k_2), ..., p_\lambda(x_T|k_T)\}$, where $p_\lambda(x|k_t)$ is the Gaussian component with maximum likelihood for frame x_t. Trim this time sequence by keeping only one component for any immediately repeating Gaussian components. Thus, we obtain a model for *sentence* X, as a time sequence of maximum-likelihood Gaussian components. Denote this sentence model, for training sentence X of speaker λ, as

$$S_{\lambda,X} = \{p_{\lambda,X}(x|s) : s = 1, 2, ..., \$_{\lambda,X}\} \quad (2)$$

where $p_{\lambda,X}(x|s)$ are Gaussian components in G_λ, sequenced by their orders of occurrence, s, to match training sentence X, and $\$_{\lambda,X}$ is the number of Gaussian components in this representation for training sentence X. The mixture weights of the GMM are not explicitly included in the sentence model (2). As will be shown later, the weights can be implied in the matching algorithms.

It may be noted that the above model $S_{\lambda,X}$ for sentence X is similar to an HMM. Indeed, each $p_{\lambda,X}(x|s)$ in the model can be viewed as an emission probability density accounting for a segment of consecutive frames with similar likelihoods, and the sequence index s can be viewed as an index for the state. With left-to-right state transitions, the model characterizes, in a statistical way, the complete temporal dynamics in X, from acoustic to lexical and to language to form the complete sentence. Traditional GMM approaches perform speaker recognition based on G_λ. In the following, we describe an approach performing recognition based on $S_{\lambda,X}$, or indirectly on G_λ through $S_{\lambda,X}$, for all the training sentences X for speaker λ. The difference between the new approach and the GMM approach is important: the GMM approach allows consecutive frames of a testing sentence to be matched by any sequence of Gaussian components in G_λ, while the new approach emphasizes the match by the Gaussian sequences forming the training sentences. The new system, thus, exploits similarities both between the short-time features and between their dependencies for recognition.

Recently, there are studies into example-based approaches to speech and speaker recognition (see, e.g., [9], [11]). These approaches seek more accurate representations of speech signals by making less assumptions about their characteristics. The above speaker-sentence model, (2), lies between the "do-nothing" model (e.g., templates) on one extreme, and the "heavy-handed" model (e.g., GMMs) on the other, representing a balance between the capture of long-range dependencies and the smoothness of the representation.

2.2. Detecting Matching Segments

In recognition, we look for similar segments of consecutive frames between the testing sentence and the sentence model, and base the decision on the degree of the similarity. We consider a full search for the similar segments. Let $O = (o_1, o_2, ..., o_N)$ be a testing sentence with N frames. To find similar segments between O and sentence model $S_{\lambda,X}$, we take every segment from O and compare it with every potential left-to-right state sequence in $S_{\lambda,X}$, which defines a segment in training sentence X to be compared with the testing segment. Denote by $o_{\tau,t} = (o_\tau, o_{\tau+1}, ..., o_t)$ a testing segment in O, consisting of consecutive frames from time τ to t, and by $s_{\tau,t} = (s_\tau, s_{\tau+1}, ..., s_t)$ a state sequence in $S_{\lambda,X}$, consisting of some continuous states forming a potential match for $o_{\tau,t}$. We measure the similarity between $o_{\tau,t}$ and $s_{\tau,t}$ by using the following probability expression:

$$P_{\lambda,X}(s_{\tau,t}|o_{\tau,t}) = \frac{p_{\lambda,X}(o_{\tau,t}|s_{\tau,t})}{\sum_{\lambda',X'\in\lambda',s'_{\tau,t}\in S_{\lambda',X'}} p_{\lambda',X'}(o_{\tau,t}|s'_{\tau,t})} \quad (3)$$

where $p_{\lambda,X}(o_{\tau,t}|s_{\tau,t})$ is the likelihood that $o_{\tau,t}$ matches state sequence $s_{\tau,t}$ in sentence model $S_{\lambda,X}$, and this is compared to the likelihoods associated with all possible state sequences that could match $o_{\tau,t}$, considering all the sentence models of all the speakers as shown in the denominator. Note that $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ has characteristics of the posterior probability of $s_{\tau,t}$ given $o_{\tau,t}$, assuming equal prior probabilities for all the state sequences. The probability $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ has an important characteristic: it favors the continuity of the matching segments, in terms of giving larger values to longer, continuous frame-state matches. To show this, rewrite $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ as a function of the individual likelihood ratios between different state sequences, i.e.,

$$P_{\lambda,X}(s_{\tau,t}|o_{\tau,t}) = 1/[1 + \sum_{\lambda',X',s'_{\tau,t}\neq s_{\tau,t}} \frac{p_{\lambda',X'}(o_{\tau,t}|s'_{\tau,t})}{p_{\lambda,X}(o_{\tau,t}|s_{\tau,t})}] \quad (4)$$

Assume that $o_{\tau,t}$ and $s_{\tau,t}$ are a pair of matching segment and state sequence, such that $p_{\lambda,X}(o_{\tau,t}|s_{\tau,t}) \geq p_{\lambda',X'}(o_{\tau,t}|s'_{\tau,t})$ for any λ', X' and $s'_{\tau,t} \neq s_{\tau,t}$. Expressing $o_{\tau,t}$ as a union of two consecutive subsegments $o_{\tau,\gamma}$ and the complement $o_{\gamma+1,t}$, where $\tau \leq \gamma \leq t-1$, we can have

$$\begin{aligned} \frac{p_{\lambda,X}(o_{\tau,t}|s_{\tau,t})}{p_{\lambda',X'}(o_{\tau,t}|s'_{\tau,t})} &= \frac{p_{\lambda,X}(o_{\tau,\gamma}|s_{\tau,\gamma})p_{\lambda,X}(o_{\gamma+1,t}|s_{\gamma+1,t})}{p_{\lambda',X'}(o_{\tau,\gamma}|s'_{\tau,\gamma})p_{\lambda',X'}(o_{\gamma+1,t}|s'_{\gamma+1,t})} \\ &\geq \frac{p_{\lambda,X}(o_{\tau,\gamma}|s_{\tau,\gamma})}{p_{\lambda',X'}(o_{\tau,\gamma}|s'_{\tau,\gamma})} \end{aligned} \quad (5)$$

The last inequality is obtained because $p_{\lambda,X}(o_{\gamma+1,t}|s_{\gamma+1,t}) \geq p_{\lambda',X'}(o_{\gamma+1,t}|s'_{\gamma+1,t})$ based on the assumption that $s_{\gamma+1,t}$ matches $o_{\gamma+1,t}$. Applying (5) to (4), we obtain

$$P_{\lambda,X}(s_{\tau,\gamma}|o_{\tau,\gamma}) \leq P_{\lambda,X}(s_{\tau,t}|o_{\tau,t}) \quad \text{for any } \tau \leq \gamma \leq t-1 \quad (6)$$

Equation (6) indicates that higher probabilities are obtained when longer segments are matched. Furthermore, we can show that higher probabilities are obtained when successive matching segments are treated as a whole than as separated segments, i.e.,

$$P_{\lambda,X}(s_{\tau,\gamma}|o_{\tau,\gamma})P_{\lambda,X}(s_{\gamma+1,t}|o_{\gamma+1,t}) \leq P_{\lambda,X}(s_{\tau,t}|o_{\tau,t}) \quad (7)$$

Inequality (7) holds due to (6) and due to $P_{\lambda,X}(s_{\gamma+1,t}|o_{\gamma+1,t}) \leq 1$. With the important properties (6) and (7), the probability function $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ can be used as a detector to detect matching segments with large continuities. Maximizing $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ among variable τ, t and X will lead to matching segments with large continuities between the testing sentence and the training sentence models. We can thus base recognition on the similarity of these large matching segments. Large segments usually contain richer and more distinct temporal dynamics. Match or mismatch of long-range temporal dynamics could be an important indication of the match or mismatch of the speakers.

Given a testing sentence $O = (o_1, o_2, ..., o_N)$, we perform a full search for the matching segments. We assume that if matched, a segment $o_{\tau,t}$ with length L ($L = t - \tau + 1$) can be accounted for by M

consecutive states $s_{i,j} = (s_i, s_{i+1}, ..., s_j)$, $M = j-i+1$, in a training sentence model. An average comparison between L and M can be estimated conveniently by examining the training sentences over their sentence models, and this is used in our system. Thus, for every $\tau \in (1, N)$ and $t > \tau$, we calculate the likelihood $p_{\lambda,X}(o_{\tau,t}|s_{\tau,t})$ using the Viterbi algorithm, where $s_{\tau,t}$ is the most-likely state sequence formed on state set $s_{i,j}$ from training sentence model $S_{\lambda,X}$. The computation is performed for every state set $s_{i,j}$ in $S_{\lambda,X}$ with $1 \leq i, j < \$_{\lambda,X}$, for every training sentence X of speaker λ. Based on the likelihoods, we form the probabilities $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ according to (3). As shown above, given a τ, $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ would increase with t if there is a continuing match between $o_{\tau,t}$ and $s_{\tau,t}$, and would start to decrease at the t where o_t is severely mismatched by s_t (assuming that it is now matched by a different state s'_t). Therefore, for each τ, we only retain the $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ up to the t before it starts to decrease. The retained $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ indicates all pairs of potentially continuously matching segments between the training and testing data, including the segments with only a single frame (i.e., $o_{\tau,t} = o_\tau$).

2.3. Algorithms for Speaker Recognition

For each speaker λ, we use dynamic programming (DP) to combine the segment probabilities $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ into an overall probability for the complete testing sentence, which is used in the recognition decision to either confirm or reject the speaker. DP is used to maximize this sentence probability by selecting the matching segments with large probabilities. This will result in a score focusing on large matching segments (since they have large segment probabilities) and hence improving the speaker discrimination. Let $\delta_\lambda(t)$ represents a partial logarithmic score ending at time t ($1 < t \leq N$) for speaker λ. We can have the following recursion:

$$\delta_\lambda(t) = \max_\tau [\delta_\lambda(\tau - 1) + L_{\tau,t} \max_{X \in \lambda, s_{\tau,t} \in S_{\lambda,X}} \ln P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})] \quad (8)$$

The maximization inside the brackets seeks the state sequence that has maximum similarity to $o_{\tau,t}$ in terms of maximum probability, and the search is performed within all the training sentences for speaker λ. In (8), the score for each segment $o_{\tau,t}$ is the segment probability $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$ weighted by the corresponding segment length $L_{\tau,t} = t - \tau + 1$. This weighting converts the segment probability $P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})$, which has a value within [0, 1] regardless of the segment length $L_{\tau,t}$ as shown in (3), to a score proportional to the corresponding segment length. Thus, the overall sentence probabilities for different segmentations become comparable, which are only a function of the length of the testing sentence, independent of the segmentation.

Further, an alternative to (8) can be obtained, by replacing the maximization over the state sequence with a summation over all the state sequences belonging to the same speaker, i.e.,

$$\delta_\lambda(t) = \max_\tau [\delta_\lambda(\tau - 1) + L_{\tau,t} \ln \sum_{X \in \lambda, s_{\tau,t} \in S_{\lambda,X}} P_{\lambda,X}(s_{\tau,t}|o_{\tau,t})] \quad (9)$$

Given a testing segment, (8) considers only the best match while (9) takes into account all potential matches from the training data. It is interesting to note that, by limiting each searched segment to a single frame, i.e., $o_{\tau,t} = o_\tau$, (9) is reduced effectively to a form of GMM. Specifically, in this single-frame segment case, each summed term in (9) is a frame-based Gaussian likelihood of speaker λ normalized by a speaker-independent constant (i.e., (3)); the sum is over all Gaussian components of the speaker through the sum over all the speaker's training sentences; different Gaussian components may appear different times in the sum depending on how often they occur in the training sentence models – this is equivalent to assigning them different weights, according to their individual importances in the training data. This is indeed a from of GMM and our new approach, thus, effectively include the GMM approach as a special case.

For a testing sentence with N frames, the score $\delta_\lambda(N)$ calculated using (8) or (9) is ready to be used for speaker identification. With a further normalization against the length of the testing sentence, the score will become usable for speaker verification. The normalized score for verification is obtained as follows:

$$\bar{\delta}_\lambda(N) = \frac{1}{N} \delta_\lambda(N) \quad (10)$$

Our approach is different from previous approaches described in [5], [9], which rely on a speech recognizer to find similar linguistic units (e.g., triphones or words) between the training and testing data. Nor our approach is similar to the ALISP approach [4] which only models and compares quasi-stationary segments in the speech data. Our new approach is completely data driven, and is capable of automatically detecting the matches of arbitrary-length, stationary or nonstationary, time processes between the training and testing sentences. The found matching processes can be linguistical, such as similar phones or phone strings, or less or non-linguistical, such as similar interjections or any sounds from the speaker which may not carry a linguistic identity but may be used to identify the speaker.

3. EXPERIMENTAL RESULTS

Experiments were conducted on the NIST 1998 Speaker Recognition Evaluation database, consisting of telephone, conversational speech data from the Switchboard-II corpus. The database contains 250 male and 250 female speakers. These speakers serve both as target speakers and as impostors. NIST has designed a set of experiments on this database, dependent on the training and testing conditions. As preliminary experiments, we considered two particular cases: each speaker is trained using two separate conversations from the same phone number, each conversation lasting about 1 min (i.e., two-session training), and the testing sentences are from the same and different phone numbers, respectively, each having a duration about 3 sec (i.e., 3s, same number/different number testing).

We trained a GMM with 128 Gaussian components as the seed model for each speaker, and built a sentence model for each training conversation on the seed model. In recognition, the algorithms described in Sections 2.2 and 2.3 were used to search matching segments between the training and testing sentences and produce the matching score. While the algorithms assume that the matching segments can have arbitrary length, from a single frame to a complete sentence, in the experiments we limited the maximum matching length to 20 frames. This reduces the amount of computation. The speech was divided into frames of 20 ms at a frame period of 10 ms. Each frame was modeled by a feature vector consisting of five subbands derived from a 25-channel mel-scale filter bank. First-order derivatives, calculated over the range of $\pm$ 2 frames, were added to the frame vector. Subband features were used in the experiments for their potential robustness to local frequency-band corruption [12].

Fig. 1 and 2 present the results by the proposed new approach, compared to a GMM using the same type of features and same number of mixtures for each speaker. The new approach offered consistent improvement, reducing the equal error rate (EER) from 11.4%

to 9.8% for the same number test, and from 29.9% to 27.1% for the different number test.

Fig. 3 shows the histogram of the length of the matching segments between the testing sentences and training sentence models decided by the new algorithm during the same number test. These multi-frame segments account for 48% of the total duration of the testing sentences. The remaining testing durations were matched with single frames selected from the models without following the training-sentence temporal dynamics. This may indicate that, over these durations, there is no significant matching dynamics between the training and testing sentences.

4. CONCLUSIONS

This paper described a new approach for modeling and matching long-range temporal dependencies in free-text speech data. The new approach uses a sentence model to represent up to sentence-level dependencies in the training data, and uses a full-search algorithm to locate the matches of arbitrary-length segments between the training and testing sentences. The sentence model is built upon a combination of statistical and example-based approaches. The search algorithm is optimized to increase the probability for the match of long, continuous segments. Preliminary experiments on the NIST 1998 SRE database have shown the potential of the new model to offer improved performance for text-independent speaker recognition.

5. REFERENCES

[1] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," ICASSP'1993, pp. 391–394.

[2] E. S. Parris, and M. J. Carey, "Discriminative phonemes for speaker identification," ICSLP'1994, pp. 1843-1846.

[3] D. Petrovska-Delacretaz, J. Cernocky, J. Hennebert, and G. Chollet, "Text-independent speaker verification using automatically labelled acoustic segments," ICSLP'1998.

[4] A. E. Hannani and D. Petrovska-Delacretaz, "Exploiting high-level information provided by ALISP in speaker recognition," Nonlinear Speech Processing Workshop, 2005, pp. 19-22.

[5] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained Gaussian mixture models," ICASSP'2002.

[6] L. Ferrer *et al.*, "Modeling duration patterns for speaker recognition," Eurospeech'2003, pp. 2017-2020.

[7] A. G. Adami1, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," ICASSP'2003.

[8] F. Farahani, P. G. Georgiou, and S. S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," ICASSP'2004, pp. 89-92.

[9] D. Gillick, S. Stafford, and B. Peskin, "Speaker detection without models," ICASSP'2005, pp. 757-760.

[10] H. Aronowitz, D. Burshtein, and A. Amir, "Text independent speaker recognition using speaker dependent word spotting," ICSLP'2004.

[11] M. De Wacher, *et al.*, "Data driven example-based continuous speech recognition," Eurospeech'2003.

[12] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," IEEE Trans. Audio, Speech, Language Process., vol. 15, pp. 1711-1723, 2007.

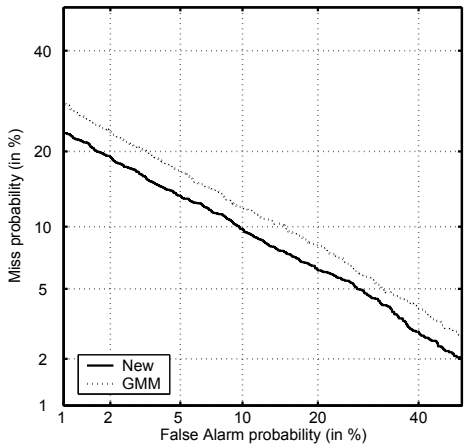

Fig. 1. DET curves for the proposed new approach and GMM for two-session training, same number test with 3-s durations.

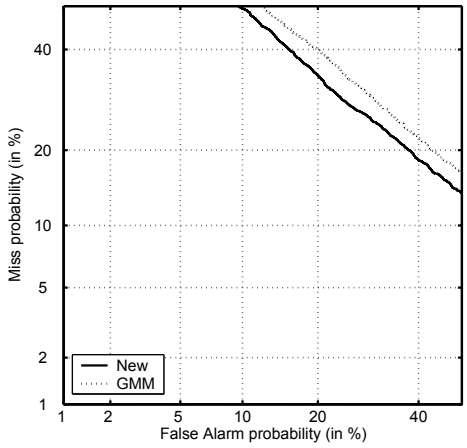

Fig. 2. DET curves for different number test with 3-s durations.

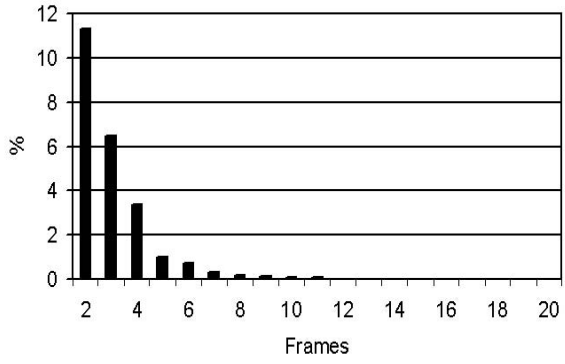

Fig. 3. Histogram of the length of the matching segments between the training and testing sentences (3s, same number test).