Decision Combination in Speech Metadata Extraction

Xiaofan Lin
Imaging Systems Laboratory
HP Laboratories Palo Alto
HPL-2003-12 (R.1)
September 22nd , 2003*

E-mail: xiaofan.lin@hp.com

speech recognition, metadata extraction, decision combination, multi-layer perceptron, gender classification

Speech metadata extraction can both improve speech recognition and enable novel Interactive Voice Response applications. Unlike the previous research, which concentrates on the frame-level signal processing and pattern classification, this paper systematically studies the behavior of decision combination at the utterance level. We analyze the asymptotic characteristics, and the factors affecting frame-level classification. In addition, we introduce new methods to more accurately and efficiently combine frame-level decisions, including phoneme/power-based weighting and smart sampling. Experimental results in gender classification are presented.

* Internal Accession Date Only Approved for External Publication
To be published in and presented at the 37th Asilomar Conference on Signals, Systems and Computers, 9-12 November 2003, Pacific Grove, CA

Decision Combination in Speech Metadata Extraction

Xiaofan Lin
Hewlett-Packard Laboratories, 1501 Page Mill Road, MS 1203
Palo Alto, CA 94304, USA
Email: xiaofan.lin@hp.com

***Abstract*-Speech metadata extraction can both improve speech recognition and enable novel Interactive Voice Response applications. Unlike the previous research, which concentrates on the frame-level signal processing and pattern classification, this paper systematically studies the behavior of decision combination at the utterance level. We analyze the asymptotic characteristics, and the factors affecting frame-level classification. In addition, we introduce new methods to more accurately and efficiently combine frame-level decisions, including phoneme/power-based weighting and smart sampling. Experimental results in gender classification are presented.**

I. Introduction

Speech metadata extraction refers to the process of intelligently inferring information about the speaker from speech signals. The types of information can be the gender [1][13], language [2], accent/dialect [3][4], age [5], and identity of the speaker [6][8]. Speech metadata extraction can directly benefit automatic speech recognition (ASR) because acoustic and/or linguistic models can be refined for different types of speakers to offer better recognition [1]. In addition, novel Interactive Voice Response (IVR) systems can be directly built on top of speech metadata extraction. For example, customized services can be offered to different groups of users.

Although some methods treat the utterance of a sentence or paragraph as a whole and directly make the final decision, many algorithms base the utterance-level result on the frame-level recognition. Fig. 1 shows the workflow of those algorithms. First, the signals of the utterance are divided into frames, which may overlap each other. It is a common practice to have the frames spaced at around 10 milliseconds with a width of 20 to 30 milliseconds. Second, each frame is individually processed. Features such as Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) [9] are extracted from each frame. Sometimes derivative features are also calculated by utilizing the features from neighboring frames (for example, the delta and delta-delta MFCC features). Then pattern classification is conducted based on those features. Various classifiers such as Gaussian Mixture Model (GMM) [6], Multi-layer Perceptron (MLP) [13] and Support Vector Machine (SVM) [8] can be applied to this step. Third, the results from individual frames are combined to make the final decision.

Most existing research in this field is concerned with the first two steps: how to improve the signal processing to get more reliable features and how to build more accurate classifiers. Simple averaging or voting across all of the frames is usually used for the last step. In contrast, this paper attempts to analyze the combination step in depth and proposes better methods to combine frame-level decisions. In fact, there are many interesting questions on the combination side: How many frames are needed? What kind of information can benefit the combination? How can we design better combination functions? How can we reduce the computation load? This paper attempts to answer these questions.

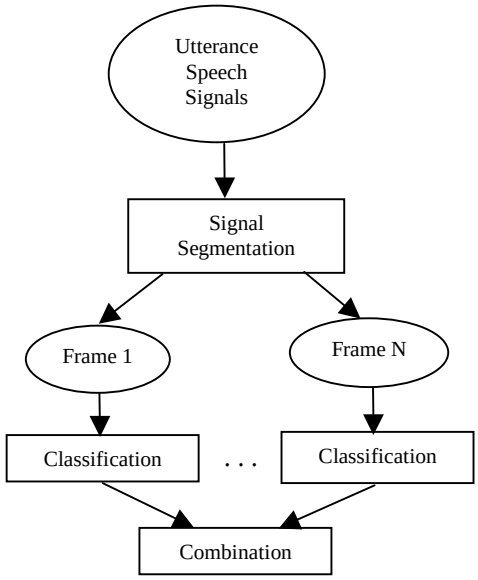

Fig. 1. Decision combination in speech metadata extraction

II. Information Levels Of Combination

As Xu, et al, proposed in [10], combination can be carried out on three levels. The abstract level combination involves only the label of the best choice. The rank level combination utilizes the ranking information of candidates. The measurement level combination attempts to achieve better combination by taking into the confidence scores of individual results. Understandably, proper use of confidence scores will potentially lead to results no worse those

achievable on the other two levels. Now let us examine the situation of speech metadata extraction.

A. Experimental System

Throughout this paper, we use gender classification to demonstrate the experimental results. The goal is to classify the speaker of an utterance as either male or female. Each utterance is first evenly segmented at an interval of 10 milliseconds. The size of each frame is 25.6 milliseconds. Then the 13 MFCC features are extracted from each frame using the algorithm implemented in Carnegie Mellon University's Sphinx 2.0 Speech Recognition System [11]. The first component only reflects the power density and thus is not useful in frame-level classification. The remaining 12 features are fed into a 3-layer MLP classifier with a 20-node hidden layer. The training section of Linguistic Data Consortium's (LDC) TIDIGITS (connected digits) database [12] is used to train the classifier, and the testing section (about 250,000 frames or 1,200 utterances) is used for evaluation.

The frame-level recognition rate is 70.43% over the frames in the testing section. This performance is in line with published results [13]. Fortunately, we have hundreds of frames in a typical utterance and it is the collective behavior of those individual results that decides the final utterance-level performance.

B. Combinations on Different Levels

Using simple majority voting for the combination, we have achieved utterance-level recognition rate of 93.43%.

Since a MLP neural net outputs continuous numbers ranging between 0 and 1 at the output nodes, we can do the confidence-level combination. The output values are first normalized. Then the following decision function is adopted:

$$D_i = \sum_{j=1}^{M} O_{ij}, \text{ where } O_{ij} \text{ is the normalized output of MLP network} \quad (1)$$

and M is the number of frames in an utterance, i is the class label.

Using (1), we have achieved utterance-level recognition rate of 95.43%. The use of the neural net outputs has reduced the error rate by 30% relatively. The difference is not difficult to understand. When only the label information is used, the classifier can only vote 1 or 0 for a class even if it is not confident about the results. When confidence scores are introduced, the classifier can express its opinion more precisely, for example, favoring "female" to "male" by 0.6 vs. 0.4. Consequently, the final decision becomes more accurate.

III. Asymptotic Analysis

In previous section, we have shown that the recognition rate can be dramatically increased from 70.43% on the frame level to 95.43% on the utterance level through combination over hundreds of frames (On average there are 208 frames per utterance in the testing data set). The next question we would like to answer is: How many frames are necessary to achieve reasonably good combination results? That is the motivation behind the asymptotic analysis.

Pi is defined as the recognition rate under the condition that the first i frames of each utterance are used. Fig. 2 displays how Pi is affected by i. One point worth mentioning is that the curve starts at about 50% instead of the frame-level 70.43%. That is because the first few frames in most utterances are just silence, which does not carry any distinguishing information. The recognition rate increases dramatically during the first 70 frames. Then it levels off.

This asymptotic characteristic is desirable in real-time applications. For example, IVR systems want to know the gender of the speaker as soon as possible to decide the emphasis of services. ASR engines also want to apply gender-dependent models from the very beginning.

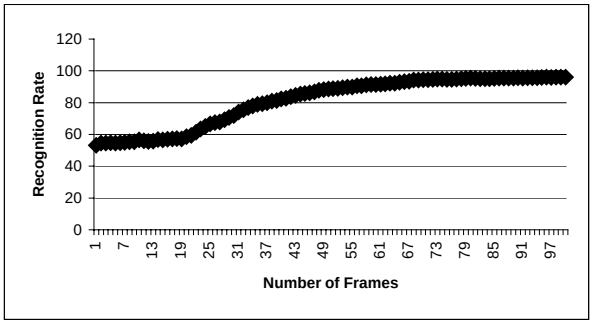

Fig. 2. Relationship between number of frames and recognition rate

IV. Frame Weighting

In the previous sections, all of the frames in the same utterance are treated equally. However, some frames are more important than the others. If those critical frames are assigned larger weights, the utterance-level decision can be more accurate. The challenge is to decide what parameters are good indicators of individual frames' importance. In this paper, two parameters are identified as reliable indicators.

A. Phonemes

Phonetic research shows that characteristics of pronouncing phonemes dictate different speaking styles or dialects [14]. Generally speaking, vowels play a bigger role than consonants. Our experiments have confirmed this observation. Sphinx V2.0 ASR is used to group frames into phonemes. Then we count the percentage of correctly recognized frames for each phoneme Since TIDIGITS only contains utterances of digits, we need to count only the 20 phonemes encountered in digits. It can be seen from Fig. 3.that frames in vowels generally enjoy higher recognition rate.

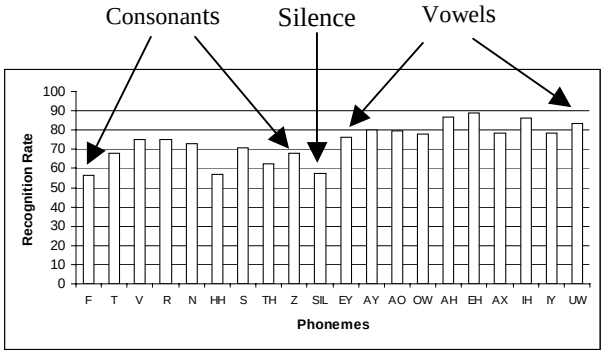

Fig. 3. Recognition rate as function of phonemes

B. Power Density

One drawback associated with phonemes is that it takes a full-fledged ASR to accurately extract phonemes, and the system's efficiency will be significantly compromised. Many real-world applications expect the metadata extraction to be a lightweight component running in the front end. So it is desired to find other indicators that can be more easily computed. The power density of each frame proves to be an ideal parameter in this regard. Conveniently, the first component of the 13 MFCC features reflects the absolute power density. In order to compensate for the volume variance among utterances, we adopt the relative difference ("PowerDiff") between the maximal power density within an utterance and a frame's absolute power density. To facilitate statistics, the continuous values are rounded off to the nearest integers. The recognition rates for different power densities are drawn in Fig. 4. For frames with very small PowerDiff, the frame-level recognition rate is around 88%. The rate drops to about 50% on frames with large PowerDiff. So we can see that the power density also serves as a good indicator of frame-level reliability.

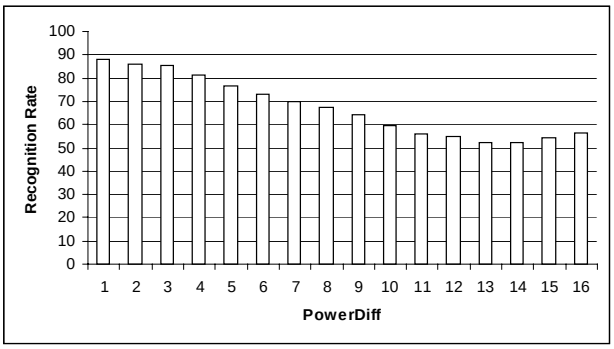

Fig. 4. Recognition rate as function of PowerDiff

C. Frame Weighting

Based on the analysis of previous subsections, the following equation is used to reflect the different roles played by individual frames:

$$D_i = \sum_{j=1}^{M} \text{Weight}(\text{Indicator}(X_j))O_{ij}, \text{ where } O_{ij} \text{ is the normalized output of MLP network} \quad (2)$$

and M is the number of frames in an utterance, i is the class label.

For each frame Xj, we first find out the indicator correlated to the reliability. It can be the phoneme label, power density, or another parameter. Then the indicator is mapped to a weight that is used in the combination. The key issue is how to map indicators to weights.

For an indicator with Q possible discrete values, Q weights are to be decided for each of them to maximize the utterance-level accuracy. A greedy algorithm is designed to do the optimization. As with any greedy algorithm, the initial condition can affect the final results, which can be locally optimal. So the optimization has been done under different initial conditions and the one leading to the highest accuracy is picked. When phonemes are used as indicators, the optimal weights are show in Fig. 5 and the resulting utterance-level recognition rate is 97.2%. When power densities are used, the optimal weights are show in Fig. 6 and the resulting utterance-level recognition rate is 97.8%. So in the two cases, the error rate is reduced by about 40% and 50% relatively compared with that using equal weights.

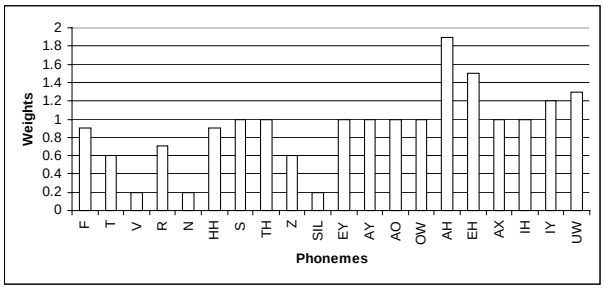

Fig. 5. Weights for different phonemes

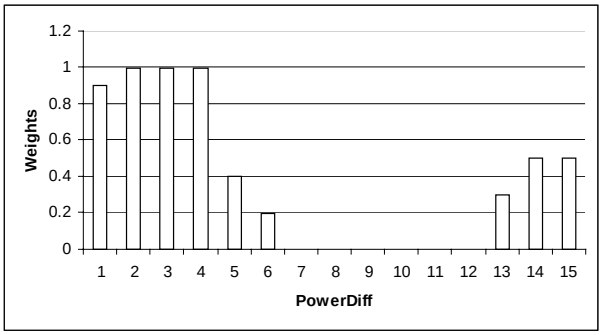

Fig. 6. Weights for different PowerDiff values

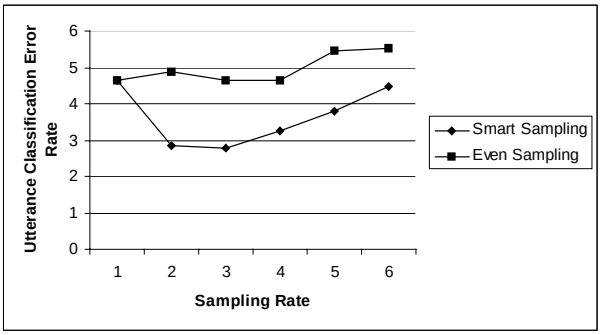

Fig. 7. Comparison of smart sampling and even sampling

V. SMART SAMPLING

One finding of the previous section is that high-power-density frames play a more important role in the combination. When the weights of low-power-density frames are set to 0, those frames can be discarded altogether. Accordingly, the time spent on extracting metadata can be reduced. Consequently, we can do "Smart Sampling" in speech metadata extraction. For each frame in the utterance, a fitness function (for example, F(X)=-PowerDiff(X)) is calculated. If the output is below the given threshold TH, the frame is discarded. Otherwise, metadata are extracted from the frame and participate in the final combination. The value of TH controls the sampling rate. Basically, Smart Sampling is a fast algorithm of frame weighting.

To demonstrate the effectiveness of the smart sampling, we have compared it with even sampling, in which the selected frames are equally spaced. Except for the sampling schemes, the other parts of the two tests are the same (Sampled frames are all assigned a weight of 1.0 and confidence scores are used). The results are shown in Fig. 7. The X-axis is the sampling rate. For example, a sampling rate of 5 means one of five frames will be selected. The Y-axis reflects the utterance-level error rates. Obviously, under the same sampling rate, smart sampling achieves lower error rate than even sampling. More interestingly, we can also see a "sweet spot" of smart sampling: When the sampling rate is between 2 and 3, the error rate is lower than when all of the frames are classified (sampling rate of 1). The error rate is actually reduced by 40% (from 4.57% to 2.8%) relatively while the speed is increased by 3 times.

Table I summarizes different combination strategies discussed in this paper. Although the error rate of smart sampling is a little higher than that achieved through power-based weighting, the low computation cost makes it an ideal candidate for real-time applications.

VI. CONCLUSIONS

In this paper, we have demonstrated that combination using confidence scores can significantly reduce the error rate. We have also analyzed the asymptotic characteristic. We have correlated the frame-level decision reliability with other factors such as phonemes and power densities. Based on the correlation, weighting methods are introduced to further improve the combination. In addition, we have proposed a Smart Sampling technique, which is able to boost the recognition rate while reducing the computation cost. One future direction is to extend the work to other tasks such as accent/dialect classification, language identification, and speaker identity detection.

ACKNOWLEDGMENT

This research is inspired and championed by HP's OpenCall Business Unit (OCBU). The author would like to thank colleagues in Intelligent Enterprise Technologies Lab of HP Labs for their valuable comments. Thanks also go to Mosur Ravishankar, whose expertise on the Sphinx system is very helpful to the implementation. The author is grateful to Pedro Moreno and Beth Logan, who have offered a lot of support and advice.

REFERENCES

[1] Y. Konig and N. Morgan, "Supervised and Unsupervised Clustering of the Speaker Space for Connectionist Speech Recognition," *Proc. of ICASSP*, Minneapolis, Minnesota, 1993.

[2] L. F. Lamel and J. L. Gauvain, "Language Identification Using Phone-based Acoustic Likelihoods," *Proc. of ICASSP*, Adelaide, Australia, 1994.

[3] M. Lincoln, S. Cox, and S. Ringland, "A Comparison of Two Unsupervised Approaches to Accent Identification," *Proc. ICSLP*, Sydney, Australia, 1998.

[4] D. R. Miller and J. Trischitta, "Statistically Dialect Classification Based on Mean Phonetic Features," *Proc. ICSLP*, vol 4, pp. 2025-2027, Philadelphia, USA, 1996.

[5] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic Estimation of One's Age with His/Her Speech Based Upon Acoustic Modeling Techniques of Speakers," *Proc. of ICASSP*, Orlando, Florida, 2002.
[6] T. F. Quatieri, *Speech Signal Processing — Principles and Practices*, Prentice Hall PTR, 2002.
[7] E. S. Parra and M. J. Carey, "Language Independent Gender Identification," *Proc. of ICASSP*, 1996.
[8] S. Fine, J. Navratil, and R. A. Gopinath, "A Hybrid GMM/SVM Approach to Speaker Identification," *Proc. ICASSP*, pp. 417-420, Salt Lake City, 2001.
[9] X. D. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR. 2001.
[10] L. Xu, A. Krzyzak, and C. Y. Suen, "Method of combining multiple classifiers and their application to handwriting recognition," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 22 no. 5, pp. 418-435, 1992.
[11] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX SYSTEM*, Kluwer Academic Publishers, Boston, 1989.
[12] Linguistic Data Consortium, LDC Database, *http://www.ldc.upenn.edu/Catalog*.
[13] Y. Konig, N. Morgan, and C. Chandra, "GDNN: A Gender-Dependent Neural Network for Continuous Speech Recognition," *International Computer Science Institute Technical Report TR-91-071*, 1991.
[14] W. Labov, S. Ash, and C. Boberg, "A National Map of the Regional Dialects of American English," *http://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html*.

TABLE I
COMPARISON OF DIFFERENT COMBINATION STRATEGIES

No	**1**	**2**	**3**	**4**	**5**	**6**
Confidence	No	Yes	Yes	Yes	Yes	Yes
Weights	Equal	Equal	Optimized on Phonemes	Optimized on Power Densities	Equal for Sampled Frames	Equal for Sampled Frames
Sampling	No	No	No	No	1:3 (Even)	1:3 (Smart Sampling)
Utterance-level Error Rate (%)	6.57	4.57	2.8	2.2	4.7	2.8