AN EFFICIENT LABELING TOOL FOR THE QUICKSIG SPEECH DATABASE

Matti Karjalainen *Toomas Altosaar* *Miikka Huttunen*

Helsinki University of Technology,
Laboratory of Acoustics and Audio Signal Processing,
Otakaari 5A, P.O. Box 3000, FIN-02015 HUT, Finland
http://www.acoustics.hut.fi/

ABSTRACT

An automated speech signal labeling tool, developed for the QuickSig speech database environment, is described. It is based primarily on the use of neural networks as diphone event detectors. For robustness, only coarse categories of diphones, such as stop–vowel and vowel–nasal, are used. 64 such detectors are implemented to cover all of the Finnish diphones. The preprocessing of speech signals is carried out using warped linear prediction and the diphone events from neural network outputs are matched to the given text transcription using a simple rule-based parser. In the case of isolated word labeling of single speaker signals a well trained system makes about 1-2 % of coarse labeling errors and the deviation of boundary positions, compared to careful manual labeling, is on average about 10 ms. Generalization ability to label other speakers shows promising.

1. INTRODUCTION

The labeling of speech signals is an important task in creating speech databases which are to be of use for other applications. E.g., phonetic analysis of a given language / dialect / speaker or the training of a speech recognizer normally presupposes the availability of labeled (time-aligned transcription) speech data.

The labeling of some given speech signal data, assuming that the orthographic or phonemic/phonetic transcription is given, can be done manually, semiautomatically, or automatically. Manual labeling is in principle the most precise and reliable method but brings about several fundamental problems. Since such work is extremely laborious and intensive, it cannot be applied to large amounts of speech data. Also, it is prone to errors; both systematic labeling biases and lack of concentration introduce inaccuracies for boundary locations. The latter problem is avoided when using automatic labeling algorithms.

If careful labeling without errors and with precise boundary locations is required, no existing automatic labeler is acceptable in practice. Thus, semiautomatic labeling systems are needed where the remaining inaccuracies from automatic labeling are corrected manually.

A typical automatic or semiautomatic system for labeling or transcription alignment is based on Hidden Markov Models (HMM) [1]. Also, the development of such a system is usually a bootstrap process where a small set of speech samples is manually labeled and an automatic labeler is trained based on this initial material. Later on the automated labeler is used to process large sets of speech data.

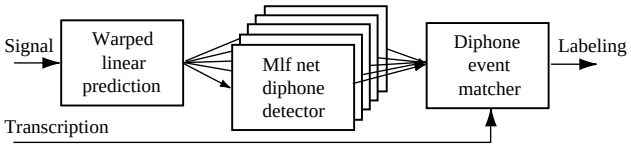

Figure 1: *Block diagram of the automated labeling system.*

In this paper we describe a new principle of automated labeling that is under development for the QuickSig speech database system [2, 3]. It is based primarily on the use of neural networks as diphone event detectors, warped linear prediction (WLP) as a preprocessing stage to compute the inputs of the networks, and a rule-based parser for matching the given transcription and the diphone event sequence from diphone detectors. The labeler shows very good time alignment precision and a low level of coarse labeling errors in a word labeling task where the system is bootstrapped by a subset of a given speech data set and tested on the remaining part of the data.

2. LABELING PRINCIPLE

Figure 1 shows the block diagram of the labeling system developed in this study. The preprocessing of speech signals could be carried out using any method that is known to work, e.g., in speech recognition. We have adopted warped linear prediction due to the reasons explained below. The preprocessed representation is applied to a set of neural networks that perform diphone event detection. Each individual net in the set is specialized to detect a specific class of diphones. The network outputs yield estimates of diphone class memberships as functions of time. Finally the diphone events, as maxima of the membership functions, are collected together and a rule-based algorithm carries out matching to the given orthographic transcription, thus yielding the desired labeling.

The QuickSig platform supports also graphical displays and interactive means for exploring and manipulating signals, transcriptions, and labeling information [2, 3, 4].

3. WLP PREPROCESSING

We have selected *Warped Linear Prediction* (WLP) [5, 6] as a preprocessor to represent signals as sequences of feature vectors. Warped linear prediction is a modification of the ordinary LP in order to implement the warped frequency scale (Bark scale) of human auditory perception. The basic idea is to replace unit delays by first-order allpass filters, i.e., frequency-dependent

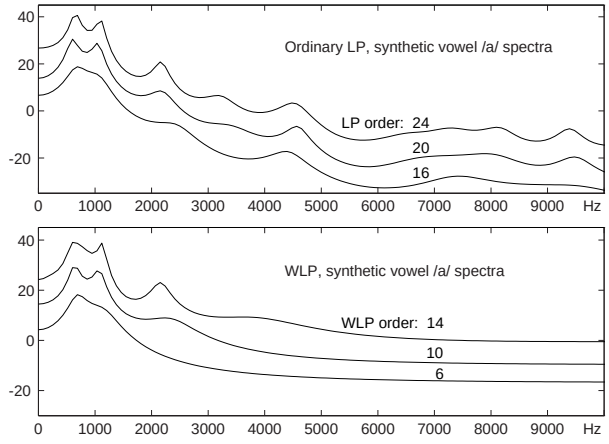

Figure 2: *LP and WLP spectra of vowel /a/ for different filter orders.*

delays, in any DSP structure, in order to obtain a warped version of it. When in linear prediction analysis the autocorrelation coefficients are computed using a warped delay line, this automatically leads to warped linear prediction.

WLP has been compared to other preprocessing methods [6, 7] and it is found to be as compact and powerful a representation as mel-cepstral coefficients (MCC). A lattice formulation of WLP with reflection coefficient parameters as outputs has a further advantage: the coefficients are normalized to lie in the range of [-1, +1]. This normalization is advantageous in our case since these parameters are used as inputs to neural networks.

Due to the Bark scale frequency warping the WLP method is a compact representation also for wide-band speech. The sampling frequency used in our speech database is 22.05 kHz. A WLP filter size of 11 was found sufficient and one more element, the signal level (loudness estimate), was added to compose a feature vector.

Figure 2 shows an example of ordinary vs. warped LP spectra for vowel /a/ for different filter orders. From the point of view of auditory resolution (Bark scale), much lower WLP orders can be used than with ordinary LP, since auditory resolution does not have to resolve spectral details at high frequencies.

4. NEURAL NET DIPHONE DETECTION

The most essential part of the labeler system is a set of diphone event detectors composed of multilayer feedforward neural nets (multilayer perceptrons). Several basic ideas are used here. First, *specialization* is applied in the form of a parallel set of neural nets, each one trained to detect a specific class of diphones. In many contexts we have found that it is better to use several simple nets, each net for a subtask, than one large network that has to solve the entire problem.

Secondly, the detectors are designed to be not too categorical so that they do not fully resolve the detailed diphone classes. Instead, *coarse categories* are used for the Finnish language so that all pair-wise combinations of {*vowel, stop, nasal, fricative, semivowel, tremulant, liquid, pause*} are provided with individual neural nets for the corresponding diphone event detection; in total 64 networks are used. This coarse-categorical analysis results in increased robustness and less sensitivity to speech and speaker variation.

The inputs to the diphone detector networks are composed of preprocessed feature vectors as shown in Fig. 3. A temporal window of ± 100 ms around the event detection point is utilized and a hop size of 10 milliseconds specifies the temporal resolution. The idea of using diphone detectors is the same as in our earlier speech recognition experiments [8]. The dimensions of each network are: 84 input nodes, 10 hidden nodes, and a single output node. Although 64 such networks are run in parallel, the computation is faster than real-time on a fast Power Macintosh which is the platform for the QuickSig system.

Figure 4 shows some examples of neural net diphone detector outputs for the word /yyteri/. The outputs can be interpreted as coarse diphone class membership estimates, 0.0 for no membership and 1.0 for full membership. During the training phase the networks learn a target membership curve that peaks around the hand-labeled phoneme boundary, being a smooth 'bump' of 25 ms and zero elsewhere. During detection, a three-point median filter is applied to smooth the network output waveforms.

Each network contributes its diphone detections that are described as discrete events of the corresponding diphone category, time position, and prominence value (peak level). A simple masking rule is used to reduce the number of low prominence events by deleting them in the vicinity of high-prominence events. In a majority of cases the correct type of event is found as the most prominent one and almost always the correct event is among the three top-prominence events.

5. TRANSCRIPTION MATCHING

The matching of a given orthographic transcription to a diphone event sequence is carried out using a relatively simple rule-based algorithm. It is based on event processing with prominence estimates and consists of three main phases:

1. First an event sequence is obtained using neural networks as described above and the events are matched to diphones in the given transcription. As a result each diphone contains a list of all potential diphone events including a prominence measure. No check of temporal positions of events is carried out yet.
2. The second step is to find all possible diphone pairs (triphones) for each phoneme generated from the orthographic transcription. This means that the previous and next di-

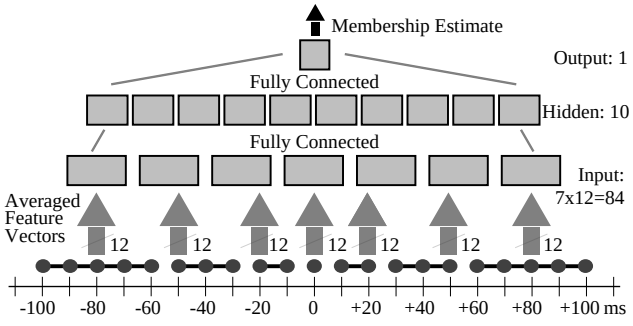

Figure 3: *The configuration of a single diphone detector neural net.*

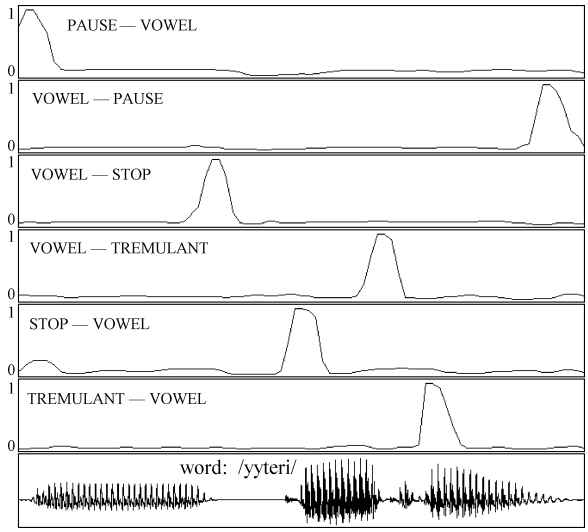

Figure 4: *Examples of diphone detector network outputs for the word /yyteri/.*

phones of the phoneme are searched for to find diphone events that properly enclose the phoneme. The combined prominence of this diphone pair is computed from the prominences of the events and their temporal distance compared with the desired duration of the phoneme. Notice that this can utilize explicit timing information. Simple averages of short and long phoneme durations are used in the present version but more detailed rule-based or neural network based duration generation could be used to improve the performance.

3. The third phase of event parsing is to check the diphones again in order to combine the triphones in such a way that they compose a consistent sequence of diphones. A list of such possible events is computed for each diphone with a combined prominence measure. If no triphone match is found, the diphone match information is used insted. This may happen when no event for a neighbouring diphone exists. In such case, also the diphone with no proper events is given a computed event that has the best rule-based approximation of temporal position between the neighbouring diphones with proper events. As a further rule, if the position of a diphone, especially inside a diphthong, deviates radically from a rule-based one a correction rule is applied to balance the position. Since the estimation of non-existing events may lead to less accurate positioning, this option is not used in the experiments below which means that there will be missing phoneme boundaries that we call coarse labeling errors. In the final phase of our algorithm the most prominent event is selected to represent each diphone in the utterance to be labeled.

6. EXPERIMENTS AND PERFORMANCE

Manually labeled speech from the Finnbet speech database [3] was used to train the neural networks and to evaluate the performance of the labeler. The speech data was high-quality 16 bit 22 kHz sampling rate recordings made in an anechoic chamber. The diphone nets were trained in all cases using standard backpropagation algorithm except that selective training was applied where the frequency of applying backpropagation was proportional to the error magnitude. In the first experiment the automated labeling tool was trained for isolated word labeling using 700 words from a single male speaker and 188 words were left for independent testing. The diphone nets were trained by applying the training material 200 times, i.e., each word and each 10 ms time position to all nets along with target data based on hand-labeling. When the networks had been trained, a testing phase followed. The 188 words were applied and the automatic labelings were analyzed by comparing with manual ones. The following table shows the percentage of coarse labeling errors and deviation of phoneme boundaries. Coarse errors are cases where the labeler did not find any diphone event to match or the category was not correct. Alignment deviations are given as the mean of absolute value of the difference in milliseconds.

Error type / data set	Test set	Train set
Coarse errors [%]	2.0	0.4
Mean align. dev. [ms]	8.7	6.2

The average deviation of the boundaries from manual segmentation was surprisingly low. Figure 5 shows the distribution of boundary errors for the test set. The result shows two facts. First, the manual segmentation has been systematic in order to allow the networks to learn, and second, the networks learn accurately. In fact, in some cases the deviation between automatic and hand-labeling turned out after closer inspection to be due to inconsistency of hand labeling.

In order to have a reference for the performance achieved, hidden Markov model (HMM) approach of simple alignment using the HTK Toolkit [1] was applied to the same data as above. An MFCC front-end with 25 ms frame and 10 ms hop size was used. For short phonemes 3 state and for long phonemes 10 state 8 mixture continuous density models were trained. The accuracy of phoneme boundary alignment was tested in two cases: training without (I) and with (II) manual segmentation information:

Train mode / data set	Test set	Train set
Mean dev. I [ms]	17.6	19.6
Mean dev. II [ms]	18.9	15.6

No coarse errors resulted since the HMMs always yield a phoneme boundary. The accuracy was not essentially different if manual segmentation data was used or not. As a conclusion of comparison, our new labeler shows an improvement in phoneme

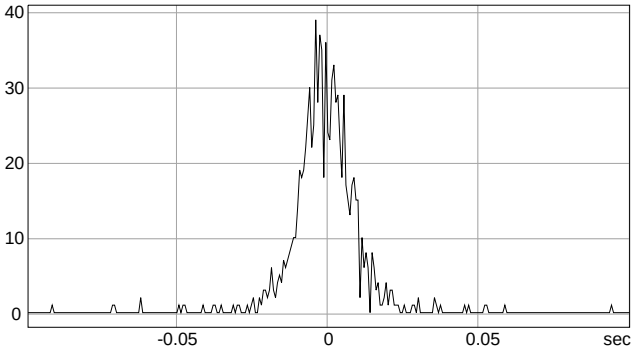

Figure 5: *Histogram of phoneme boundary deviations between automatic and hand labeling.*

boundary alignment over simple HMM techniques. Although the diphone event detectors are applied only every 10 ms the interpolation of event positions can yeld better resolution of phoneme boundaries. (Notice that since this means a resolution better than a typical pitch period, the hand-made labeling may not have such accuracy even conceptually.)

The next experiment was related to the important question of how the labeler can generalize to manage with speech data from different speakers. An interesting case was to apply the labeler trained above from speaker MK to words of another male speaker MV that were not used for training (case 1). 700 words of both speakers were used then for training in case 2 and tested with the the rest for each speaker. The results are given in the following table.

Error type / data set	Case1	Case2 MK	Case2 MV
Coarse errors [%]	12.6	7.6	9.0
Mean align. dev. [ms]	12.1	1.2	1.9

It can ce seen that when using only one speaker in training, the rate of coarse errors increases relatively much for other speakers but the alignment accuracy for the boundaries found is still good. If two speakers are used in training, the system performs well for both of them. We can conclude that the approach shows potential for speaker generalization that is important when doing labeling speaker independently.

Experiments with larger units than isolated words have to be carried out yet because properly hand-labeled material was not available. It can be expected that the performance of the labeler may drop slightly from the level of isolated words since the variation of speech parameters is larger. Otherwise the principle used should not be critically dependent on the length of the utterance to be labeled.

7. DISCUSSION

Among problems that we found in the current system is the detection of certain diphone events, such as slow transition diphones inside diphthongs. A neural net with wider temporal input frame and focus to slow transitions could improve the performance. Also vowel-liquid (/l/) transitions are often found problematic.

There is space for much improvement also in the rule-based parsing of events to diphones of given transcription. The rules desribed above are quite simple ad hoc rules and a more systematic matching algorithm could improve the accuracy. It might be worth of considering the application of an HMM-like formalism to the sequence of events found by neural nets.

The computational efficiency of the system is good. The database system runs on Power Macintosh computers and the time taken to obtain a labeling result on a 300 MHz machine is about the same as the duration of the speech signal to be labeled.

8. SUMMARY AND FUTURE WORK

This paper describes an automated speech labeling tool that is a part of the QuickSig speech database system. The labeler is based on using neural networks for finding diphone events related to coarse categories of Finnish speech and a rule-based parser to match a given orthographic transcription to a given speech signal.

The system performs with a low error rate and precise phoneme boundary assignment when applied to speech samples of a speaker that has been trained for the event detector neural nets. Since the system is based on robust coarse category features, it could be possible to extend it to labeling of speech also in a speaker-independent manner. This and other improvements of the labeler remain to be done as future work.

9. ACKNOWLEDGEMENT

This project has been supported by the Academy of Finland.

10. REFERENCES

[1] HTK Toolkit, Entropic Research Laboratory, Inc., `http://www.entropic.com/`.

[2] Karjalainen M., and Altosaar T., "An Object-Oriented Database for Speech Processing," *Proc. of Eurospeech'93*, Berlin, 1993.

[3] Altosaar T., Karjalainen M., and Vainio M., "A Multi-Lingual Phonetic Representation and Analysis System for Different Speech Databases," *ICSLP'96*, Philadelphia, 1996.

[4] Karjalainen M., "DSP Software Integration by Object-Oriented Programming, A Case Study of QuickSig," *IEEE ASSP Magazine*, April 1990.

[5] Strube H. W., "Linear Prediction on a Warped Frequency Scale," *J. Acoust. Soc. Am.*, vol. 68, no. 4 (1980), pp. 1071-1076.

[6] Laine U. K., Karjalainen M., Altosaar T., "Warped Linear Prediction (WLP) in Speech and Audio Processing," *Proc. IEEE ICASSP -94*, Adelaide, 1994.

[7] Boda P., *Psychoacoustical Considerations in Speech Analysis and Recognition*. Licentiate thesis, Helsinki University of Technology, Espoo, Finland, 1995.

[8] Altosaar T., and Karjalainen M., "Diphone-Based Speech Recognition Using Time-Event Neural Networks," *Proc. ICSLP'92*, Banff, 1992.