https://doi.org/10.3176/lu.1998.3.10

MATTI KARJALAINEN, TOOMAS ALTOSAAR (Espoo)

A TOOL FOR AUTOMATIC LABELLING OF FINNISH SPEECH

The labelling of speech signals is an important task for 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 recogniser normally presupposes the availability of labelled (time-aligned transcription) speech data. The labelling of some given speech signal data, assuming that the orthographic or phonetic transcription is given, can be done manually, semiautomatically, or automatically. Manual labelling 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 labelling biases and lack of concentration introduce inaccuracies for boundary locations. The latter problem is avoided when using automatic labelling algorithms.

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

A typical automatic or semiautomatic system for labelling or transcription matching is based on Hidden Markov Models. Also, the development of such a system is usually a bootstrap process where a small set of speech samples is manually labelled and an automatic labeller is trained based on this initial material. Later on, the automated labeller is used to process large sets of speech data.

In this paper we describe a new principle of automated labelling that is developed for the QuickSig speech database system (Karjalainen 1990; Karjalainen, Altosaar 1993). It is based primarily on the use of neural networks as diphone event detectors, warped linear prediction (WLP) as preprocessing 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 labeller shows very good time alignment precision and a low level of coarse labelling errors in a word labelling task where the system is bootstrapped by a subset of a given speech data set and tested on the remaining part of the data.

Labelling method

Figure 1 shows the block diagram of the labelling 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 pre-

diction 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 specialised to detect a specific class of diphones. The network outputs yield estimates of diphone class memberships as functions of time.

Finally, the diphone events are collected together and a rule-based algorithm carries out matching to the given orthographic transcription, and thus the desired labelling is obtained.

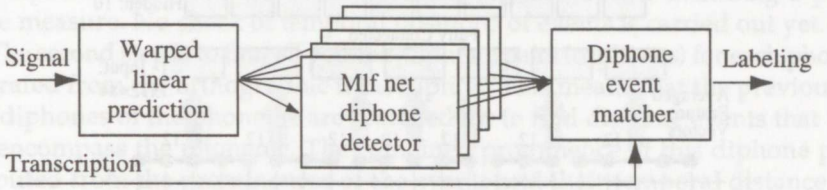

Figure 1. **Block diagram of the automated labelling system.**

The QuickSig system includes graphical displays and interactive means for exploring and manipulating signals, transcriptions, and labelling information (Karjalainen 1990; Karjalainen, Altosaar 1993).

Preprocessing by warped linear prediction

We have selected Warped Linear Prediction (WLP) (Strube 1980; Laine, Karjalainen, Altosaar 1994) 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 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 (Laine, Karjalainen, Altosaar 1994; Boda 1995) 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 normalised to lie in the range of (–1, +1). This normalisation 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.

Diphone detection by neural networks

The most essential part of the labeller system is a set of diphone event detectors composed of multilayer feedforward neural nets (multilayer perceptrons). Several basic ideas are used here. First, s p e c i a l i s a t i o n 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 one for a subtask, than one large network that has to solve the entire problem.

Secondly, the detectors are designed to be not too categoric so that they do not fully resolve the detailed diphone classes. Instead, c o a r s e c a t e g o r i e s are used for the Finnish language so that all pair-wise combinations of {vowel, stop, nasal, frica-

tive, 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 variation.

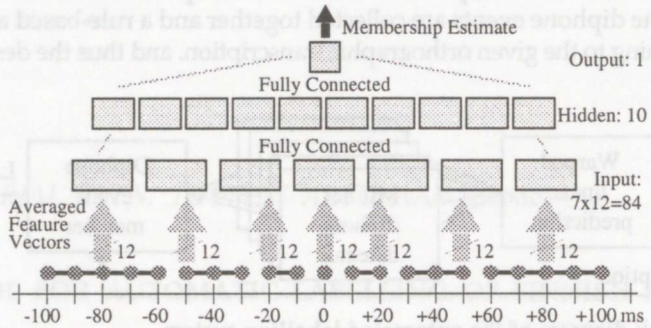

Figure 2. **The configuration of a single diphone detector neural net.**

The inputs to the diphone detector networks are composed of preprocessed feature vectors as shown in Figure 2. A temporal window of 100 ms around the event detection point is utilised 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 (Altosaar, Karjalainen 1992). 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.

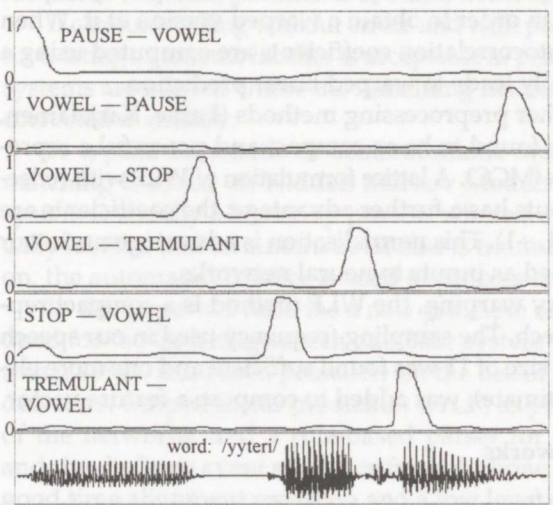

Figure 3. **Examples of diphone detector network outputs for the word /yyteri/.**

Figure 3 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-labelled 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 (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.

Matching the transcription

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:

First an event sequence is searched for 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.

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 diphones of the phoneme are searched for to find diphone events that properly encompass the phoneme. The combined prominence of this diphone pair is computed from the prominences of the events and their temporal distance compared to the desired duration of the phoneme. Notice that this can utilise 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.

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 and the most prominent event is selected to represent the diphone under study.

Labeller performance

We have tested the present version of the automated labelling tool by training the system for word labelling using 700 words from a single male speaker; 188 words were left for independent testing. The diphone nets were trained by a standard backpropagation algorithm 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-labelling.

When the networks had been trained, the testing phase followed. The 188 words were applied and the automatic labellings were analysed. It was found that 4.5% of coarse labels experienced problems: 3.1% deletions and 1.4% replacements. Due to the principle used, no insertion errors are possible. For the training set samples there were 2.5% coarse errors, all of them being deletion errors.

The average deviation of the boundaries from manual segmentation was surprisingly low: the average of absolute deviation was 7.9 ms and standard deviation 12 ms. Figure 4 shows the distribution of the phoneme boundary deviations.

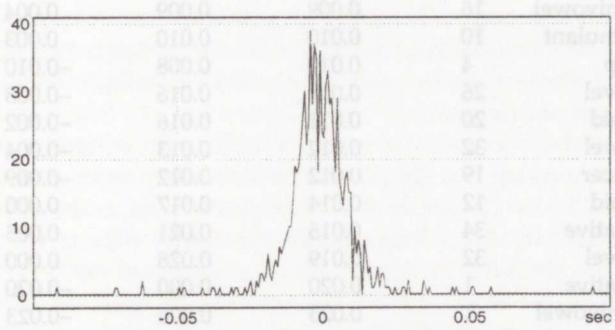

Figure 4. **Histogram of phoneme boundary deviations between automatic and hand labelling.**

The result shows two facts. First, the manual segmentation has been systematic in order to allow the networks to learn, and second, the networks perform accurately. In fact, in some cases the deviation between automatic and hand-labelling turned out after closer inspection to be due to inconsistency of hand labelling.

Many improvements are still possible in the labelling system. When analysing the errors we found that most of them are systematic and can be eliminated, e.g., by post-processing or by improving the matching rules. Systematic deletions just mean that a new diphone must be inserted and the position can be computed using duration information. In a similar way most of replacements can be post-corrected. A reduction of coarse errors should be possible such that the remaining errors are not more than about 1—2%.

Among problems that we found in the current system is the detection of diphone events when the transition is slow, such as diphthongs (vowel transitions) and vowel-liquid transitions. The performance of the former could be improved by using a neural net where the input context covers a wider temporal frame than for most of the other diphone types. Table 1 shows the error decomposition according to network class.

Table 1

Error decomposition of the 188 evaluation words sorted according to the average absolute error (AvgAbsErr) in seconds (N indicates the number of occurrences found in a specific coarse-category 37 out of a possible 64 coarse-categories existed in the word set)

Order	Left	Right	N	AvgAbsErr	Std Dev	Avg Err
1	nasal	stop	2	0.003	0.001	–0.003
2	fricative	stop	3	0.003	0.004	–0.003
3	tremulant	nasal	1	0.004	0.000	–0.004
4	vowel	stop	61	0.005	0.006	0.001
5	spacer	stop	83	0.005	0.006	–0.001
6	stop	vowel	144	0.005	0.006	–0.002
7	nasal	vowel	35	0.006	0.011	0.002
8	vowel	tremulant	20	0.006	0.006	–0.003
9	fricative	vowel	53	0.006	0.007	–0.002
10	vowel	fricative	27	0.006	0.007	–0.003
11	stop	nasal	2	0.006	0.001	–0.006
12	vowel	spacer	163	0.006	0.008	–0.001
13	stop	stop	1	0.006	0.000	0.006
14	fricative	liquid	3	0.007	0.006	–0.004
15	vowel	semivowel	14	0.007	0.009	–0.002
16	liquid	vowel	36	0.007	0.009	–0.003
17	nasal	fricative	4	0.007	0.007	–0.000
18	spacer	vowel	13	0.007	0.010	0.002
19	stop	fricative	1	0.008	0.000	0.008
20	liquid	stop	3	0.008	0.002	–0.008
21	spacer	nasal	14	0.008	0.010	0.004
22	vowel	nasal	40	0.008	0.013	0.003
23	spacer	semivowel	16	0.009	0.009	0.004
24	spacer	tremulant	10	0.010	0.010	0.003
25	tremulant	stop	4	0.010	0.008	–0.010
26	tremulant	vowel	26	0.011	0.016	–0.003
27	vowel	liquid	20	0.011	0.016	–0.002
28	semivowel	vowel	32	0.012	0.013	–0.004
29	nasal	spacer	19	0.012	0.012	–0.009
30	spacer	liquid	12	0.014	0.017	0.000
31	spacer	fricative	34	0.015	0.021	0.005
32	vowel	vowel	32	0.019	0.028	0.000
33	tremulant	fricative	1	0.020	0.000	–0.020
34	fricative	semivowel	1	0.023	0.000	–0.023
35	liquid	nasal	1	0.048	0.000	–0.048
36	fricative	spacer	4	0.050	0.018	–0.050
37	stop	semivowel	1	0.071	0.000	–0.071

No direct comparison of the present system was possible with any other labelling system. In an informal comparison we found, however, that the HTK Toolkit, when used for time alignment, typically yielded position errors of around 20 ms.

Summary and future work

This paper has described an automated speech labelling tool that is part of the QuickSig speech database system. The labeller 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 labelling of speech also in a speaker-independent manner. This and other improvements of the labeller remain as future work.

REFERENCES

A l t o s a a r, T., K a r j a l a i n e n, M. 1992, Diphone-Based Speech Recognition Using Time-Event Neural Networks. — Proceedings of ICSLP'92, Banff.

A l t o s a a r, T., K a r j a l a i n e n, M., V a i n i o, M. 1996, A Multi-Lingual Phonetic Representation and Analysis System for Different Speech Databases. — Proceedings of ICSLP'96, Philadelphia.

B o d a, P. 1995, Psychoacoustical Considerations in Speech Analysis and Recognition, Espoo (Licentiate Thesis, Helsinki University of Technology).

K a r j a l a i n e n, M. 1990, DSP Software Integration by Object-Oriented Programming. A Case Study of QuickSig. — IEEE ASSP Magazine.

K a r j a l a i n e n, M., A l t o s a a r, T. 1993, An Object-Oriented Database for Speech Processing. — Proceedings of Eurospeech'93, Berlin.

L a i n e, U. K., K a r j a l a i n e n, M., A l t o s a a r T. 1994, Warped Linear Prediction (WLP) in Speech and Audio Processing. — Proceedings of IEEE ICASSP'94, Adelaide.

S t r u b e, H. W. 1980, Linear Prediction on a Warped Frequency Scale. — Journal of the Acoustical Society of America, 68, 1071—1076.