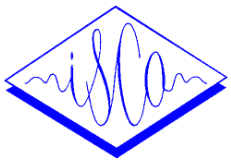

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

THE USE OF DISCRIMINANT NEURAL NETWORKS IN THE INTEGRATION OF ACOUSTIC CUES FOR VOICING INTO A CONTINUOUS-WORD RECOGNITION SYSTEM

Claude Lefèbvre and Dariusz A. Zwierzyński

Speech Research Centre, National Research Council of Canada
Building U-61, Montreal Road,
Ottawa, Ontario, Canada K1A 0R6

Abstract

The performance of a small vocabulary speaker-dependent robust speech recogniser can be improved by adding more input features in the front-end. Our present speech recognition system employs both static & dynamic spectral representations which are combined with a linear discriminant analysis. We have done recognition experiments with CVC words, differing in their initial consonant phonemes only, e.g. *peep* vs *beep* and found that most of the errors are due to the system not distinguishing between voiceless/voiced stop consonants. There are a number of acoustic cues useful to improve distinction between voiceless/voiced plosives, specifically, the fundamental frequency at voicing onset and the Voice Onset Time (VOT). This paper reports on recognition experiments where both of these features are extracted from the speech signal and are combined with the other features using the linear discriminant network. The results from the experiments confirmed that the addition of these two input features improved the performance of the recogniser for confusable word-pairs.

Introduction

In the past years, we have done speech recognition experiments with a system using linear discriminant analysis to combine various input features. The main objective was to compare this technique with conventional feature extractors like cepstral analysis. Tests demonstrated that the systems' performance in recognition of quiet and particularly of distorted speech was a lot better using the linear discriminant technique, for example, continuous-word recognition tests for tilted speech produced by male speakers resulted in low error rates of the order of 0.1% for the spectral discriminant analysis developed in our Laboratory, in comparison with a 30% error rate for the weighted cosine transform [1].

The objective of this new work is to test the use of multiple input features combined with a linear discriminant analysis for a confusable set of words differing in the initial plosive (stop) consonants. Our current system uses linear discriminant analysis to combine the static and dynamic spectral features of the speech signal. Similar features have been used in various speech recognition systems [2,3]. Recognition experiments show a fairly small substitution rate, around 7%, of errorful decisions for word-initial voiceless /p, t, k/ and voiced /b, d, g/ plosives, embedded in minimal word-pairs differing only in their initial consonants, e.g. *keep/geep* or the "ee" letter set of the alphabet, e.g. *b, d, g, p, t.* We believe that this weakness can be overcome by extracting new information from the subphonemic level of the plosive consonants and integrating it with the other two spectral representations. We specifically concentrate on such invariant cues to voicing/voicelessness in stops as voice onset time (VOT), or the distinct pitch variations in vowels following voiceless/voiced plosive consonants. A perceptual study done with words differing in their initial consonants has shown that these cues are used to make such a distinction [4]. Particularly, the VOT is a strong speech feature used to distinguish between voiceless/voiced stop consonants. The VOT is also a cue that increases the distinction of stop consonants differing in their place of articulation [5]. We believe that the integration of pitch information would help in the recognition of speech embedded in noise.

Speech recognition experiments.

The speech recognition experiments employed a set of discriminant functions, which were applied to the extracted features in the front-end component, and a dynamic time warping (DTW) technique in the back-end component (Fig. 1). The spectral features were extracted in the front-end processing stage by a simulated mel-scale spectral filter-bank based on a conventional fast-fourier-transform analysis. The representations obtained from the front-end analysis were subsequently processed by a linear discriminant network. The computation of the linear transformation involves the between-class with the within-class covariance information of

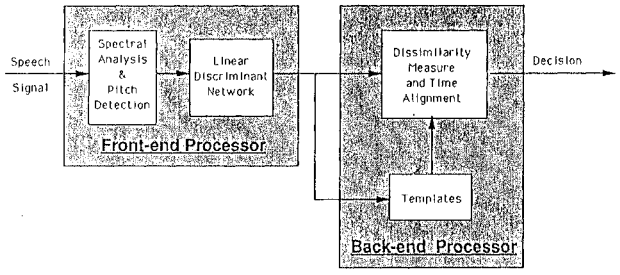

Fig. 1. Block diagram of the automatic speech recognition system used in the experiments.

the spectral features [6]. The weights in the linear discriminant network have been calculated for each speaker. This technique is very efficient at extracting invariant features for reliable speech recognition, especially when the speech is distorted or spoken in high background noise. The linear discriminant network is called IMELDA, which stands for Integrated MEL-scale Linear Discriminant Analysis, because it combines various mel-scale spectral representations into a single set of discriminant functions.

As shown in Figure 2, we have used various input features for the recognition experiments. We called the input features IMELDA-2 representations because we are combining static and dynamic spectral features of the speech signal. The results from speech recognition tests with static and dynamic spectral features serve as a basis (see Table 2). Tests were done with the addition of VOT representations which we called IMELDA-3. Recognition tests with pitch measured at voicing onset have also been done.

The speech database used in the recognition experiments consists of 5 examples of 6 CVC words differing in their initial consonants spoken by 7 male speakers. The word-pairs were *peep* vs *beep, teep* vs *deep* and *geep* vs *keep.* The average fundamental frequency measured for each speaker varied between 90 to 160 Hz and was fairly constant across the CVCs. For the recognition experiments, white noise of two different intensities has been added to the speech signal. In one case the signal-to-noise ratio was 15 dB and in the other case it was 9 dB. The spectral balance of the speech signal was modified by increasing the higher frequency by 6 dB/octave. As shown in Table 1, we have derived a confusion matrix from the recognition experiments of the CVCs using static and dynamic spectral features for all the conditions described above. Other confusion matrices were derived from the use of IMELDA-1 representations and a cosine transform [7]. All the experiments have shown that a good percentage of the errors is due to the wrong

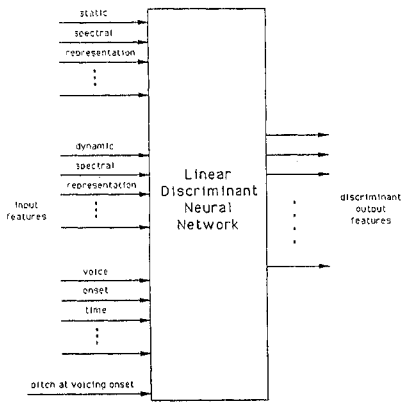

Fig. 2. Schematic representation of the linear discriminant (neural) network with input and output features. The weights in the network were computed from the spectral and microprosodic features of the speech signal.

Confusion matrix of CVCs							
Spoken word	Recognized word						
	peep	teep	keep	beep	deep	geep	Errors (%)
peep	115	15	3	4	1	2	17.9
teep	3	129	5	1	1	1	7.9
keep	0	3	122	0	0	15	12.9
beep	7	4	1	107	17	4	23.6
deep	0	12	3	9	109	7	22.1
geep	0	3	16	0	0	121	13.6

Table 1. Isolated word recognition confusion matrix using static and dynamic spectral features of 5 CVCs in a set of 6 word-pairs spoken by 7 male speakers. The test material was presented in four conditions: undegraded, with white noise added to give a 15 and 9 dB SNR and with a 6 dB/octave spectral tilt applied.

voiceless/voiced distinction. The confusion matrix derived from the static and dynamic spectral features shows that there is room for improvement in recognising velar stops.

Addition of VOT representations (IMELDA-3)

The time from stop consonant release to the beginning of the vocal cord vibration is called Voice Onset Time (VOT). The VOT is longer for voiceless stop consonants than for voiced stop consonants. The average time that separates the two groups is approximately 44 msec. VOT also increases when the place of articulation is higher in the vocal tract [5]. We can simply measure the VOT by applying twice a linear regression analysis on the static spectral representation. The static spectral features are obtained by running an FFT analysis on every 6.4 msec portion of the speech signal. The dynamic spectral features are obtained by running a linear regression analysis on 7 frames of 6.4 msec. Just as in a matched filter, the same linear regression analysis applied twice on 7 frames of the static spectral features would respond effectively to separate the two groups of VOT (see Figure 3). Recognition results using the IMELDA-3 representations are shown in Table 2. As expected, the results are better for IMELDA-3 compared to IMELDA-2 representations.

Measuring pitch at voicing onset

There has been comparatively little research into including prosodic features in the context of word recognition. There seems, however, to be considerable potential in employing pitch information in the recognition process [8,9].

Two distinctly different pitch micropatterns are observable in stressed vowels following voiceless and voiced plosive consonants, e.g. *p/b* in *peep/beep.* The pitch at the voicing onset of vowels in stressed syllables starts higher after voiceless plosives than after voiced ones [4]. Those microprosodic pitch perturbations are induced by different articulatory strategies used in the production of plosive con-

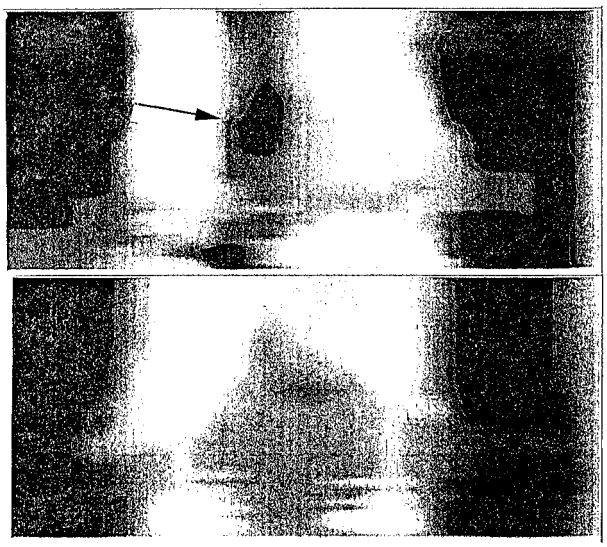

Fig. 3. VOT spectral representations obtained from the second derivative of the short-time static spectral features of the words *peep* (on top) and *beep* (on bottom). A strong peak just before voicing onset (arrow) is present in the VOT spectrogram of the voiceless stop consonant.

sonants, and comprise an invariant feature of speech manifested across different speakers and languages. It is thus possible to use that information to help determine the voicing status of the word-initial consonant preceding a stressed vowel, and to enhance the recogniser in its classification of the consonant. Detection of pitch for the CVC words performed in our laboratory has shown that the pitch value during approximately the first 30-50 msec following the first glottal closure instant can be a relevant cue for differentiating between voiceless and voiced consonants (Fig. 4).

Part of our present work focusses on the development and refinement of a reliable pitch detection algorithm. The principal premise in designing the pitch detector is that it should work for speech employing different voice qualities and be robust in noise.

Speaker-Dependent Male Isolated CVC Recognition Percentage Errors (%)				
Representation	Quiet	Noise-1	Noise-2	Tilt
IMELDA-2	4.3	11.4	29.6	23.4
IMELDA-2 & pitch	3.8	11.0	20.5	22.4
IMELDA-3	2.4	10.0	20.5	25.7
IMELDA-3 & pitch	2.9	7.6	20.5	27.6

Table 2. Isolated consonant-vowel recognition results for five examples of 6 word-pairs (e.g. *peep* vs. *beep*) spoken by 7 male speakers. The test material was presented in four conditions: undegraded (Quiet), with white noise added to give a 15 dB (Noise-1) and 9 dB SNR (Noise-2), and with a 6 dB/octave spectral tilt applied (Tilt).

The pitch detector developed in our laboratory [10] uses the following time-domain algorithms:

1) Inverse filtering,
2) Negative half-wave rectification of the LPC residual,
3) Centre-clipping,
4) Average Magnitude Difference Function (AMDF),
5) Histogram calculation for pitch determination at voicing onset.

Inverse filtering is first performed to remove the vocal tract acoustic information (i.e. formant structure) imposed on the excitation source, thus leaving distinct excitation peaks with little activity areas between them. The inverse-filtered signal is subsequently half-wave rectified and centre-clipped to enhance the peaks, and finally processed through the Average Magnitude Difference Function (AMDF).

The AMDF algorithm is an autocorrelation method that measures periodicity in a speech waveform by looking for minima [11]. Pitch tracking is done by building, for a short-time segment of speech, a histogram of possible detected pitch periods which correspond to AMDF minima. When building the histogram, we also add credence to the minimum that has other minima at its double, triple, and quadruple multiples.

First, the average pitch is estimated for a window of a given length placed at the beginning of the utterance. Once the average pitch has been estimated, the pitch for each sample in that window is calculated through a backtracking process which takes into account the detected average pitch value. Knowing the average pitch for a particular speaker enables the algorithm to reduce the search range, and therefore have a more precise measure of the actual pitch. The pitch in the following windows is calculated on the basis of the average pitch which is constantly updated as the window advances.

In our research on microprosodic phenomena, we have found that pitch tracking at vowel voicing onset can be a difficult task, since the algorithm has to work in different contexts

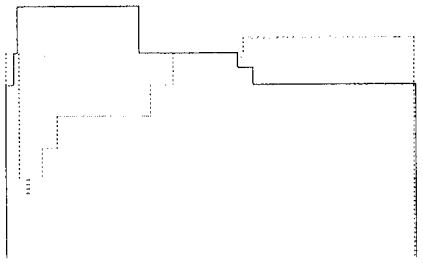

Fig. 4. Linearly time-aligned median pitch contours from five examples of the words *keep* (continuous line) and *geep* (dotted line) spoken by one speaker. Both pitch contours start at the voicing onset. Note that the pitch contour just after onset is higher for the voiceless *k(eep)* than the voiced *g(eep)* plosive consonant.

(e.g prevoicing murmur, pitch jitter). However, extensive tests of the current pitch detection algorithm on the 2100 words of the CVC MIT database and on a small database containing nasalised speech, indicate a high accuracy rate (99%) in pitch measurement.

Recognition results from the combination of pitch measured at voicing onset with the IMELDA-2 representations are shown in Table 2, as well as those from the combination of pitch with IMELDA-3 representations. The results of the tests of IMELDA-2 with pitch are much better than the tests with the IMELDA-2 representations alone. The results indicate that there is almost no improvement by adding pitch to IMELDA-3 representations compared to the IMELDA-3 representations alone.

Conclusions

Speech recognition experiments with a small speech database of word-pairs differing in their initial consonants have been performed to find out whether the addition of more input features to the static and dynamic spectral representations improves the recognition rate. As predicted, recognition results with a confusable set of words indicate that a better recognition rate can be obtained by including microprosodic features and adding more segmental speech features. We strongly believe that a higher recognition rate could be obtained with spectral representations derived from discriminant filters than from linear regression filters.

The results in Table 2 show that in general better results are obtained with the addition of pitch and VOT features. The biggest improvement for quiet speech is obtained for IMELDA-3 compared with IMELDA-2 representations. With white noise added to the speech (SNR = 15 dB), there is a significant improvement for IMELDA-3 and pitch compared to all other three representations. This seems to indicate that the glottal excitation pulses are more resistant to noise than the burst onset of the stop consonant. The test with the tilted speech might not be a valuable one because a change in the spectral balance might change the perception of stop consonants. We think that the tilted CVCs should be classified in a perceptual test done with listeners.

Various ways can be used to improve the performance of a recognition system to distinguish between voiceless/voiced stop consonants. The use of VOT calculated from the second derivative of the static spectral features is appropriate for a dynamic time warping based speech recogniser. *Deng et al.* [12] have shown that modelling accurately the stop consonants from burst to voicing onset with HMM models improves the performance of their recogniser.

In the future, we intend to do the same recognition tests with the confusable word-pairs spoken by females. In a small hardware speech recogniser using one DSP chip in the front-end stage, it would be preferable to use only VOT representations since the computation of the pitch at voicing onset is very expensive. However, cheaper computation algorithms like multi-layer perceptrons could be envisaged for measuring pitch.

Acknowledgments

We thank Carl Swail for helpful comments on a preliminary version of this paper. We are grateful to the NRC/IAR Flight Research Laboratory for allowing us to use the computing facilities. The work on a further development of the NRC ASR system constitutes a part of a larger project realised in co-operation with the R&D Department of the Neil Squire Foundation, Vancouver, Canada.

References

1. Lefèbvre, C. and Starks D., Small Vocabulary Continuous Speech Recognition in a Helicopter Cockpit, *Proc. Military and Government Speech Tech '89,* pp. 36-41, Arlington, November, 1989.
2. Aikawa, K. and Furui, S., Spectral Movement Function and its Application to Speech Recognition, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-88),* New York, pp. 223-226, April 1988.
3. Hunt M. J. and Richardson S. M., Use of Linear Discriminant Analysis in a Speech Recogniser, *Proc. Voice Systems Worldwide,* London, May 1990.
4. Silverman, K., F_0 Segmental Cues Depend on Intonation: The Case of the Rise After Voiced Stops, *Phonetica,* Vol. 43, pp. 76-91, 1986.
5. O'Shaughnessey D., *Speech Communication, Human and Machine* Addison-Wesley Publishing Company, 1987.
6. Hunt, M. J. and Lefèbvre, C., Distance Measures for Speech Recognition, *Aeronautical Note,* NAE-AN-57, Ottawa, March, 1989.
7. Lefèbvre, C. and Zwierzyński, D. A., On the Use of F_0 Variations in Automatic Speech Recognition, *J. Acoust. Soc. Am.* Vol. 87, Supl. 1, p. S105, 1990.
8. Lea, W. A., Prosodic Aids to Speech Recognition. In: W. A. Lea (Ed.) *Trends in Speech Recognition,* pp. 166-205, Englewood: Prentice-Hall, 1980.
9. Vaissière, J., The Use of Prosodic Parameters in Automatic Speech Recognition. In: H. Niemann *et al.* (Eds.) *NATO ASI Series, Recent Advances in Speech Understanding and Dialog Systems,* Vol. F46, pp. 71-99, Berlin-Heildelberg: Springer-Verlag, 1988.
10. Zwierzyński D. A. and Lefèbvre C., Improvement of the NRC Automatic Speech Recognition System, *Proc. of the Canadian Conference on Electrical and Computer Engineering* Ottawa, Canada, September 1990.
11. Fette, B., Gibson, R. and Greenwood, E., Windowing Functions for the Average Magnitude Difference Function Pitch Extractor, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-80),* pp. 49-52, Denver, April, 1980.
12. Deng, L., Lennig, M., and Mermelstein, P., Modeling Microsegments of Stop Consonants in a Hidden Markov Model Based Word Recognizer. *J. Acoust. Soc. Am.,* Vol. 87, pp. 2738-2747, June 1990.