THE EXTRACTION AND INTEGRATION OF SELECTED CUES FOR VOICING INTO A CONTINUOUS-WORD AUTOMATIC SPEECH RECOGNITION SYSTEM

Dariusz A. Zwierzyński and Claude Lefèbvre

Neil Squire Foundation and National Research Council
Speech Research Centre, Building U-61
Montreal Road, Ottawa, Ontario, Canada K1A 0R6

ABSTRACT

This paper deals with the enhancement of automatic recognition of stops in isolated word-initial stressed syllables of minimal word-pairs in normal and degraded speech. It was found that the error rate for stops could be decreased by incorporating the cues to voicelessness/voicing extracted from a short-time spectral content of the speech signal. The acoustic cues comprised: F_0 perturbations in post-occlusive vowels, voice onset time (VOT), and the presence/absence of low-frequency at voicing onset. Results confirm that additional input features improve performance.

1. INTRODUCTION

Automatic recognition of stops in word-initial position in minimal word-pairs poses problems for recognition systems. Our studies indicated that many errors were caused by the system not distinguishing between the voicing status of the initial stops. Apparently, the spectral information from the filter-bank analysis was insufficient or too weak for the system to make an efficient distinction.

We propose a method of reducing the error rate, by extracting new information on the voicing status from the subphonemic level of a stop, and integrating it with other spectral information derived by the filter-bank analysis. This concept represents a phonetically-based approach to recognition of stops, which enhances the successful performance of algorithms employing linear discriminant analysis (LDA) developed in Canada [1] and implemented in our recognition system.

2. SPEECH RECOGNITION SYSTEM

The recognition system used in the experiments is a speaker-dependent, robust, small-vocabulary, continuous speech recogniser. The system performs very well in quiet and in distorted speech [2].

Spectral representations from the filter-bank analysis are processed by a linear discriminant network. Our system can use static and dynamic spectral representations as input to the linear discriminant network [Fig. 1].

The linear discriminant analysis generates a set of discriminant functions which are applied to the acoustic features extracted from the speech signal. As the linear discriminant network combines various mel-scale spectral representations into a single set of discriminant functions, the transformations have been called IMELDA for integrated mel-scale LDA.

3. EXTRACTION OF NEW FEATURES

3.1. F_0 perturbations

The speech database used in our experiments included 5 examples of 6 CVC minimal word-pairs comprising the 6 stops and the vowel /i/ recorded by 7 male speakers.

F_0 at voicing onset of vowels after voiceless stops starts higher and falls deeper than it does after voiced stops where it either slightly falls or remains level [5]. The deep F_0 fall is one of the stronger cues to a voiceless stop.

Fundamental frequency perturbations are extracted by a modified custom-designed AMDF pitch detector, operating on the raw audio waveform and outputting pitch data from the initial frames at voicing onset. [4]. The pitch information is then integrated with the spectral information.

Experimental results demonstrate that the benchmark IMELDA-2 representation was outperformed when supplemented with pitch information (IMELDA-P) extracted from voicing onset (Tab. 1). The improvement is evident across the 4 acoustic experimental conditions.

3.2 Voice Onset Time

In English VOT is longer for voiceless stops than for voiced ones. The average time separating the two voicing categories is approx. 44 msec. Two different methods were used to derive the VOT spectral representations; in both of them the burst and VO points had been manually labelled.

In the first method, a VOT representation was obtained from a second derivative of the static spectral features (LCE). The linear regression analysis applied twice on 7 frames of 6.4 msec of LCE gives an effective separation of the two VOT groups

Table 1 illustrates that in comparison with IMELDA-2, the addition of VOT resulted in improved recognition in all 4 conditions. Compared with IMELDA-P, IMELDA-VOT is worse only in Tilt.

In another experiment IMELDA-2 was combined with pitch and VOT (Fig. 2). A comparison of IMELDA-PVOT with IMELDA-VOT (Tab. 1) reveals improvement in Noise-1, but the error rate is slightly higher for Quiet and Tilt. Yet, here again, IMELDA-PVOT outperformed IMELDA-2 in all 4 conditions.

The conclusion derived from this series of experiments is that IMELDA-VOT constitutes the optimal representation for the four acoustic conditions. The addition of pitch and VOT to IMELDA-2 indicates that improvement is possible, however it is not so uniform as with IMELDA-VOT and IMELDA-P, and better than IMELDA-VOT only in one experimental acoustic condition. With white noise added (15dB SNR) there is an improvement with IMELDA-PVOT over the other representations, which may suggest that in this condition the glottal excitation pulses are more resistant to noise than the burst pulse. Indeed, better recognition in noise was obtained with all representations incorporating the extracted cues. We suspect that tilting the spectral balance resulted in higher error rates with IMELDA-VOT and IMELDA-PVOT, as by changing the gain in each channel, the perceptual salience of the voicing cues may thereby have been changed.

Deriving a VOT representation from a 2nd derivative is effective but too slow for real-time applications. Hence, a faster method was developed.

4. TIME-DOMAIN FILTERS

Linear discriminant time-domain filters were derived from the use of linear discriminant analysis on the LCE representation of quiet speech. (Fig. 3). The VO instant was the timing reference point. Between- and within-class onset periods were computed for 4 and 13 frames before VO and 4 and 3 frames after VO, giving a time-domain filter with 8 or 16 coefficients.

The filter was then applied to all frames of the analysed word. Also, experiments were conducted with a different number of discriminant filters. Using 8 frames enabled us to study the contribution of cues detectable in the time domain in the VO area. By using 16 frames, the discriminant filter could be derived from running LDA on the segment spanning an entire VOT area, including both the burst and the voicing onset.

In the 8 frame analysis, we wanted to extract the cues related to the F1 transition, and specifically the absence or presence of the low-frequency energy after VO which cues the perception of a voiceless/voiced stop respectively [3]. Table 1 illustrates that 3 discriminant filters (IMELDA-3f8c) produce the best results for this condition, outperforming IMELDA-2 in the 4 acoustic conditions.

In the other analysis, discriminant filters were computed over 16 frames to maximise the range by deriving filters from a VOT area, capturing the possible burst, VOT, and F1 cues in one type of analysis. The results (Tab. 1) reveal that 4 time-domain discriminant filters (IMELDA-4f16c) perform much better in quiet and with 15 dB SNR than IMELDA-2, IMELDA-VOT, and IMELDA-PVOT, and slightly worse with 9 dB SNR and in tilted speech. On the other hand, it turns out that IMELDA-3f8c outperforms IMELDA-4f16c in degraded speech, and is slightly worse in Quiet. It may be here that degraded speech obliterates the burst, and thus the VOT cue, much more than it happens when the filters are derived from the voicing onset area only.

Complete results from tests with adding pitch data to IMELDA-3f8c/4f16c were not available at the time of writing. Initial data suggest that adding pitch can improve the performance in degraded speech.

5. DISCUSSION AND CONCLUSIONS

The incorporation of additional cues to voicelessness/voicing into our recognition system has decreased the error rate for stop consonants. Despite using a small speech database, we believe that the benefits from incorporating the new input features have been demonstrated. It has to be emphasised that IMELDA and the time-domain filters were derived from a quiet speech representation. Previous findings [2] indicate, however, that deriving an IMELDA transform from integrated quiet, noisy and tilted speech representations produces better results in degraded speech, and only negligibly worse in quiet, than when the transform is derived from quiet speech only. We shall publish corresponding results obtained with IMELDA representations computed on degraded speech as they become available. We expect that computing the IMELDA transform and the time-domain filters on degraded speech may result in an improved performance of the more versatile IMELDA-4f16c compared to IMELDA-3f8c.

The usage of time-domain discriminant filters provides for focussing the analysis on different areas of interest, as shown in our experiments. Similarly, different time-domain filters could be used for processing intervocalic stops, where the VOT cue may not be reliable, or word-final stops, where again the presence and salience of the relevant cues is different. In addition, time-domain filters are suitable for real-time implementation, and will be incorporated into the future versions of the hardware recogniser jointly developed by the Neil Squire Foundation, the Canadian Marconi Company, and the National Research Council of Canada. Current work in this area also includes the refinement of a neural network (MLP) for real-time pitch and voicing onset/offset detection for the automatic calculation of VOT representations [6].

6. ACKNOWLEDGEMENTS

Part of this work was conducted under a contract (041ST.W7714-0-3529) with the DND Canada. We are grateful to Gary Birch and David Starks for discussions on various topics related to research reported in this paper.

7. REFERENCES

[1] HUNT, M. J. and LEFEBVRE, C. (1989), "Distance measures for speech recognition", *Aeronautical Note, NAE-AN-57*, Ottawa, March.

[2] HUNT, M. J. and LEFEBVRE, C. (1989), "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech", *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proc., ICASSP-89*, Glasgow, Scotland.

[3] KLATT, D. H. (1975), "Voice onset time, frication, and aspiration in word-initial consonant clusters", *J. of Speech and Hear. Res., 18*, 686-706.

[4] LEFEBVRE, C. and ZWIERZYNSKI, D. A. (1990), "On the use of F_0 variations in automatic speech recognition", *J. Acoust. Soc. Am. 87, Supl. 1*, S105.

[5] SILVERMAN, K. (1986), "F_0 segmental cues depend on intonation: The case of the rise after voiced stops", *Phonetica, 43*, 76-91.

[6] ZWIERZYNSKI D. A. and LEFEBVRE C. (1990), "Improvement of the NRC automatic speech recognition system", *Proc. of the Canadian Conf. on Electr. and Comp. Eng., 2*, 5.3.1.-5.3.4, Ottawa, Canada.

Table 1. Isolated CVC recognition results for 5 examples of 6 minimal-word pairs spoken by 7 male speakers. Test material presented in 4 conditions: undegraded (Quiet), with white noise added to give a 15dB SNR (Noise-1) and 9dB SNR (Noise-2), and with a 6dB/octave tilt applied (Tilt).

SPEAKER-DEPENDENT ISOLATED CVC RECOGNITION ERRORS (%)				
Representation	Quiet	Noise-1	Noise-2	Tilt
IMELDA-2	4.3	11.4	23.4	29.6
IMELDA-P	3.8	11.0	20.5	22.4
IMELDA-VOT	2.4	10.0	20.5	25.7
IMELDA-PVOT	2.9	7.6	20.5	27.6
IMELDA-3f8c	1.4	3.3	16.6	23.3
IMELDA-4f16c	0.0	6.2	24.8	32.4

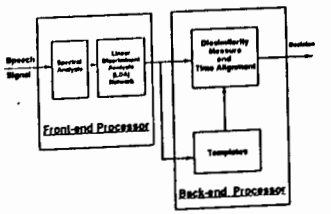

Fig. 1. Block diagram of the speech recognition system

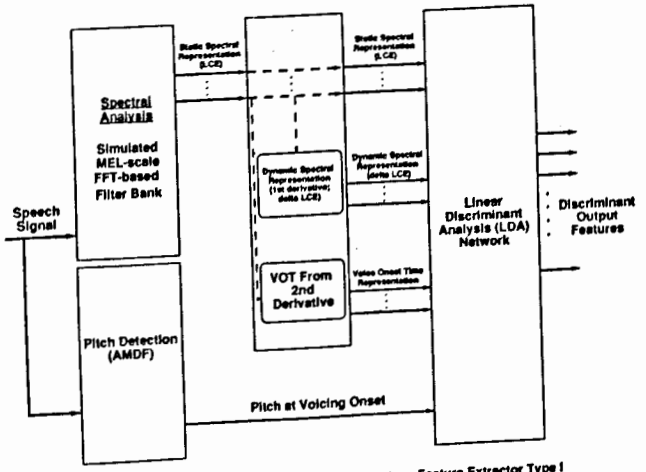

Fig. 2. Block diagram of the automatic speech recognition system: Feature Extractor Type I

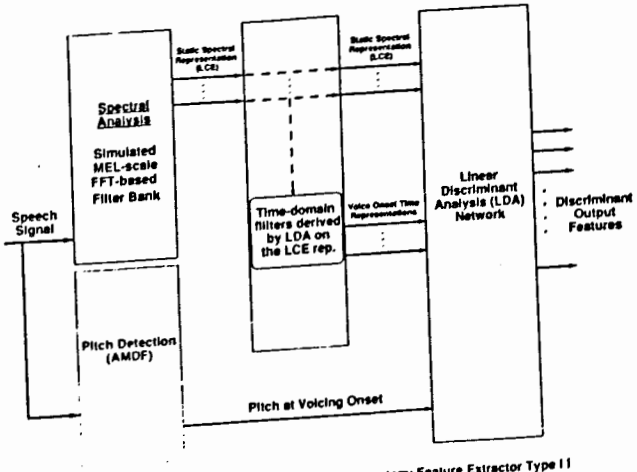

Fig. 3 Block diagram of the automatic speech recognition system: Feature Extractor Type II