

Speech separation for speech recognition

A. DE CHEVEIGNÉ, H. KAWAHARA*, K. AIKAWA* and A. LEA*

Laboratoire de Linguistique Formelle, CNRS, Université Paris 7, Case 7003, 2 place Jussieu, 75251 Paris cedex 05, France

**ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-02, Japan*

résumé: Divers modèles de séparation de voix parasites sont implémentés à l'entrée d'un système de reconnaissance de la parole. L'objectif est d'estimer, au travers des taux de reconnaissance, l'efficacité des principes de traitement. Un premier modèle utilise l'annulation harmonique des voix parasites dans le domaine temporel. Nous analysons les limites de son efficacité, et étudions les moyens de dépasser ces limites. Un deuxième modèle utilise le renforcement harmonique de la voix cible. Ce deuxième modèle se révèle moins efficace, du fait de la nature non stationnaire de la parole. Ce résultat apporte un élément de réponse à une question explorée dans un autre article présenté à ce congrès: le système auditif utilise-t-il le renforcement harmonique, l'annulation harmonique, ou les deux, pour séparer des voix simultanées?

1. INTRODUCTION

People understand speech at signal-to-noise ratios that defeat speech recognition systems. To do so, they rely on various factors, both acoustic and cognitive [1, 2, 3, 4, 5, 6]. One factor that has been studied extensively is harmonicity, as exploited when components of a mixture differ in fundamental frequency (F0) [7, 8, 9, 10, 11, 12, 13, 14]. Various models and processing methods have been proposed for harmonic sound separation (see [15] for a review). Logically, these methods have two cues to operate on: the harmonic structure of the voice to be extracted (the target), and that of the interfering voice (the ground). A companion paper presented at this congress [16] investigates whether the auditory system uses one cue in preference to the other. Here we try to find out which of the two is more useful in eliminating an interfering voice in a preprocessing stage before a speech recognition system. Recognition rate gives a convenient (if task-dependent) measure of efficacy.

We also examine in detail the factors that limit the efficacy of a model of harmonic sound cancellation based on time-domain comb filtering controlled by the period of the interfering voice. We propose a scheme by which the effects of one of these factors (spectral distortion) can be reduced.

2. METHODS

Task and database

The task was to recognize 100 target words by pattern matching to a set of 100 templates (consisting of the same words spoken by the same speaker). The target speech was corrupted by adding interfering speech (words belonging to the task set, but different from the target) at various signal-to-noise ratios (SNR). SNR was defined as a fixed factor applied to signals before mixing; the actual SNR within individual pairs could be quite different.

The database consisted of 100 short Japanese words taken from the ATR database [17]. Speech was sampled at 12 kHz, 16-bit resolution. Dynamic time warping (DTW) pattern matching [22] was performed on 16-coefficient vector arrays. Vectors were derived from 128-coefficient magnitude spectra by averaging

coefficients 8 at a time. Spectra were calculated using 256-point Hanning windows at a frame rate of 128 samples.

Formal significance tests were not performed. However, based on the criteria of the McNemar test [19], one can give an upper limit on significance of individual differences. Differences of 5 words or less do not meet a 5% significance level. Larger differences may or may not meet this level, depending on how errors are distributed.

F0 estimation and comb filtering

F0 estimation from mixed speech is a major task in speech separation [15]. Here we bypass it completely by estimating the fundamental period directly from the speech before mixing, using an algorithm similar to that described in [15, appendix B-2], without error correction or smoothing of any sort. To enhance resolution, speech was upsampled 4 times by linear interpolation before period estimation. The search range for the period was set to allow an F0 range of 60 to 300 Hz. Estimates were produced at a frame rate of 2.5 ms. Where periodicity was poor (unvoiced portions or transitions), as indicated by low values of a periodicity measure, period estimates were set to zero. Harmonic cancellation was implemented in the time domain by use of a comb filter with the following impulse response:

$$h_n = \delta_n - \delta_{n-l} \quad (1)$$

The lag l is controlled by the period estimate of the interference. Linear interpolation was used to accommodate fractionary estimates (permitted by upsampling in the estimation stage). Filtering was performed only for valid (non-zero) period estimates. The onset and offset of filtering was smoothed by panning between filtered and unfiltered signals using a 4.2 ms raised-cosine ramp. Harmonic enhancement was likewise implemented as a comb filter with the following impulse response (the order K was varied as a parameter):

$$h_n = \frac{1}{K} \sum_{i=0}^{K-1} \delta_{n+il} \quad (2)$$

3. RESULTS

3.1. Harmonic cancellation

Speech mixed with interference was first fed to the speech recognition stage without cancellation. The SNR was set to $-\infty$ (no signal), 6, 0, -6, -12 and $+\infty$ (no interference) on a decibel scale. The lower dotted line in Figure 1 shows the recognition rate. The rate is 100% for no interference (in this case the task is trivial), and 0% for no target (as interference consists of task words different from the target).

The continuous line shows rates after harmonic cancellation. Recognition is improved at low SNR, but not at high SNR. This is because cancellation introduces spectral distortion of the target that counterbalances the effect of removing a small amount of interference (if the task for noise-free conditions were less trivial, performance at the SNR= ∞ point would be worse than without cancellation). In no case does cancellation bring the rates near the 100% level that a practical system might require.

Two factors limit the effectiveness of harmonic cancellation. One is the residue that remains due to imperfect periodicity of the interfering speech. The other is the spectral distortion of the target due to the comb filter. We can separate these factors. Linearity allows us to swap the stages of mixing and comb-filtering [20]. We can then apply comb-filtering to the interference only: this isolates the effect of the cancellation residue (distance from 100% of upper dotted line in Fig. 1). We can then estimate the effect of spectral distortion by difference (distance between upper dotted line and continuous line). Relatively small at small SNR, spectral distortion is the major factor that limits the efficacy of harmonic cancellation at large SNR.

3.2. How to reduce the impact of spectral distortion?

Spectral distortion affects recognition because distorted targets don't match undistorted templates very well. A solution that comes to mind is to also adjust the templates to improve the match. The difficulty is to estimate the proper adjustment. If the target is harmonic, spectral distortion can be understood as a "moiré"

effect between the line spectrum of the target and the transfer function of the comb-filter. This can be described by the transfer function:

$$|H(f)| = |\sin(2\pi f|T - L|)|$$

where T is the period of the target, and L is the lag of the cancellation comb filter. The crosses in Fig. 1 show rates obtained after interference cancellation is combined with template adjustment. Template adjustment does not completely eliminate the effect of spectral distortion (compare with upper dotted line), and it has little effect at low SNR, but it considerably improves recognition in the high SNR region useful for applications. Whereas cancellation requires only the F0 of the interference, template adjustment requires the F0s of both target and interference, and works only when both are periodic. In our database, target and interference were voiced for 56% of all frames, but *simultaneously* voiced for only 41%.

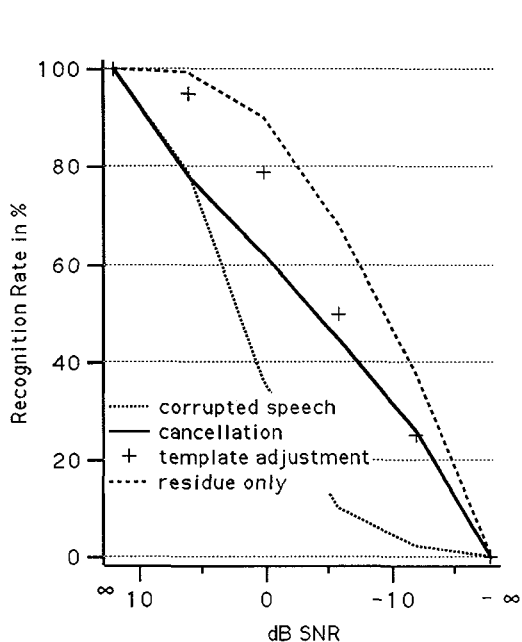


Fig. 1. Lower dotted line: recognition rates as a function of SNR for corrupted speech. Continuous line: same after interference cancellation. Upper dotted line: rates for filtered interference added to the unfiltered target. Crosses: rates obtained with template adjustment.

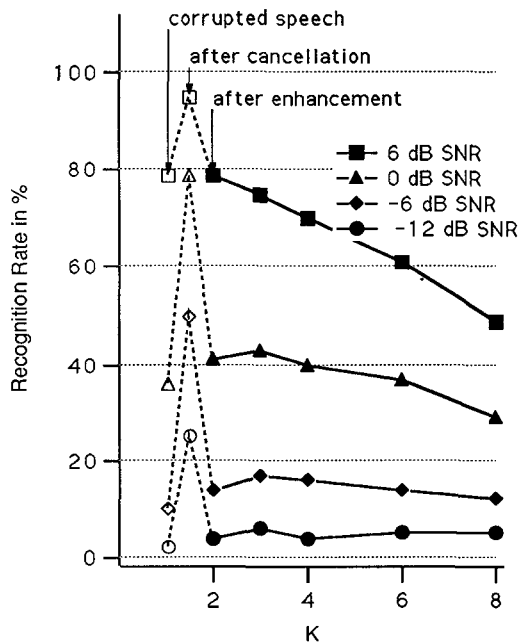


Fig. 2. Full lines: recognition rate after harmonic enhancement, as a function of the number of "prongs" in the impulse response (parameter K in Eq. (2)), for various SNR. Leftmost points correspond to unfiltered corrupted speech, and harmonic cancellation with template adjustment.

3.3 Cancellation vs enhancement.

A priori, each strategy has its advantages and drawbacks:

- Cancellation works whatever the target (voiced or unvoiced), but the interference must be harmonic.
- Cancellation is easier to perform when SNR is low, because it requires estimating the F0 of the interference.
- Cancellation causes spectral distortion of the target.
- Cancellation can be implemented effectively with a filter with a short impulse response.
- Enhancement works whatever the interference, but the target must be harmonic (only voiced parts can be enhanced).
- Enhancement is easier to perform when SNR is high because it requires estimating the F0 of the target.
- Enhancement causes no spectral distortion if the target is perfectly harmonic.
- Effective enhancement requires a filter with a long impulse response [15, Appendix A].

The long impulse response required by enhancement might make it impractical for speech, which is non-stationary. We investigated this question. Harmonic enhancement was implemented with a filter determined

by Eq. (2) and evaluated with the same recognition task. The filter's lag parameter was driven by an estimate of the F0 of the target speech. Recognition rates were measured for various values of the parameter K as displayed in Fig 2, together with rates for harmonic cancellation. For all but 6 dB SNR, enhancement is most effective for K=3. It remains much less effective than harmonic cancellation.

4. CONCLUSION

Harmonic cancellation is effective for reducing interference and improving recognition rates when SNR is small. Template adjustment extends this improvement to large SNR, and may allow harmonic cancellation to be of practical use in speech recognition systems. Such systems would have to solve the task of F0 estimation that we ignored here. Overall, the improvement is equivalent to a noise reduction of 6-9 dB, but our results depend closely on the task we chose. We used simple time-domain processing, but it seems likely that results for frequency-domain schemes would follow a similar pattern.

Harmonic enhancement is considerably less effective than cancellation, probably because of the non-stationarity of speech. An adaptive version of the comb-filter [21] might perform better, but it is unlikely it would bridge the gap with cancellation. The result is of interest because it may help explain why the auditory system prefers cancellation to enhancement in tasks where both are possible a priori [16].

ACKNOWLEDGEMENTS

This research was supported by both ATR and CNRS. We wish to thank Harald Singer for sharing his implementation of the McNemar test, and Jean Laroche for signal-processing advice and comments on a previous draft.

REFERENCES

- [1] Brox, J. P. L. and S. G. Nootboom (1982), "Intonation and the perceptual separation of simultaneous voices", *J. Phonetics*, 10, 23-36.
- [2] Cherry, E. C. (1953), "Some experiments on the recognition of speech with one, and with two ears", *JASA* 25, 975-979.
- [3] Darwin, C. J. (1981), "Perceptual grouping of speech components differing in fundamental frequency and onset-time", *QJEP*, 33A, 185-207.
- [4] Darwin, C. J. and J. F. Culling (1990), "Speech perception seen through the ear", *Speech Comm.* 9, 469-475.
- [5] Bregman, A. S. (1990), *Auditory scene analysis*, MIT Press: Cambridge, Mass.
- [6] McAdams, S. (1989), "Segregation of concurrent sounds. I: Effects of frequency modulation coherence", *JASA* 86, 2148-2159.
- [7] Assmann, P. F. and Q. Summerfield (1990), "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies", *JASA* 88, 680-697.
- [8] Carlyon, R., L. Demany and C. Semal (1992), "Detection of across-frequency differences in fundamental frequency", *JASA* 91, 279-292.
- [9] Culling, J. F. and C. J. Darwin (1993), "Perceptual separation of simultaneous vowels: within and across-formant grouping by F0", *JASA* 93, 3454-3467.
- [10] Gardner, R. B., S. A. Gaskill and C. J. Darwin (1989), "Perceptual grouping of formants with static and dynamic differences in fundamental frequency", *JASA* 85, 1329-1337.
- [11] Lea, A. (1992), "Auditory models of vowel perception", Thesis, University of Nottingham.
- [12] Scheffers, M. T. M. (1983), "Sifting vowels", Thesis, University of Groningen.
- [13] Summerfield, Q. (1992), "Roles of harmonicity and coherent frequency modulation in auditory grouping", in "The auditory processing of speech", edited by B. Schouten, Mouton de Gruyter, 157-165.
- [14] Zwicker, U. T. (1984), "Auditory recognition of diotic and dichotic vowel pairs", *Speech Comm.* 3, 256-277.
- [15] de Cheveigné, A. (1993), "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing.", *JASA* 93, 3271-3290.
- [16] de Cheveigné, A., S. McAdams, J. Laroche and M. Rosenberg (1994), "Identification de voyelles simultanées harmoniques et inharmoniques", *this conference*.
- [17] Kuwabara, H., Y. Sagisaka, K. Takeda and M. Abe (1989), "Construction of ATR Japanese speech database as a research tool",
- [18] Tohkura, Y. (1987), "A weighted cepstral distance measure for speech recognition", *IEEE Trans. ASSP*, 35, 1414-1422.
- [19] Gillick, L. and S. J. Cox (1989), "Some statistical issues in the comparison of speech recognition algorithms", *IEEE ICASSP*, 532-535.
- [20] Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach.", Thesis, University of Sheffield.
- [21] Frazier, R. H., S. Samsam, L. D. Braid and A. V. Oppenheim (1976), "Enhancement of speech by adaptive filtering", *IEEE ICASSP*, 251-253.
- [22] Sakoe, H. and S. Chiba (1978), "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. ASSP* 26, 43-49.