Titre: Speech separation for speech recognition.

Auteurs:

Alain de Cheveigné
Laboratoire de Linguistique Formelle (CNRS/Université Paris 7), case 7003, 2 place Jussieu, F-75251 Paris Cédex 05, France, tel. (33-1) 44273633, fax (33-1) 44277919, email: alain@linguist.jussieu.fr

Hideki Kawahara, Kiyoaki Aikawa, Andrew Lea
ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-02, Japan, tel. (81) 7749-5-1020, fax (81) 7749-5-1008

resumé: 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: