

AUTOMATIC SPEAKER IDENTIFICATION IN C2 CENTRES: CHALLENGES AND PITFALLS

Peter J. Chatelain¹

Abstract. In future command and control (C2) centres, staff should be able to move about freely, unconstrained by microphone headsets, while their conversations are transcribed to text using speaker-independent speech-recognition devices. The output of the transcriber would be automatically labelled with the staff's identities. Automatic Speaker Identification (ASI) is a candidate to perform that back-end function. However, ASI accuracy remains lower than that of human speaker recognition, despite 40 years of R&D, because the technology does not cope well with a minority of speakers. In addition, the acoustical environment of C2 centres is very complex. It is strongly affected by reverberation and the cocktail-party and Lombard effects. These and the presence of low-energy speech degrade both speech and speaker recognition. Nevertheless, that degradation is expected to vary significantly from centre to centre. It is possible that the use of microphone arrays can render ASI operational, in at least some of those centres, especially if the number of speakers considered is small.

INTRODUCTION

In text-independent automatic speaker recognition (ASR), a computerized system identifies an individual by analysis of his/her speech. ASR comprises both verification (ASV) and identification (ASI). In ASV, an identity is claimed and the system must attempt to authenticate that claim. In ASI, the claimant is known to belong to a well-defined set of identities (closed set) or is suspected of belonging to that set, without this being known with certainty (open set). One of these technologies may be relevant to command and control (C2) centres where there is a need to label recorded utterances during conferences, briefings or planning sessions, for dissemination or archival purposes. This paper exposes the main technical hurdles which need to be overcome or adequately managed should ASI be successfully exploited in such environments.

AN IDEALIZED ASI SCENARIO

Those C2 centres considered here are medium to large indoor venues as opposed to automated mobile command posts. The centres are frequented by clerical staff and members of a well-defined pool of military decision makers. These regular staff include a few dozen people, at most. In addition, an indeterminate number of people visit the centres infrequently. Decisions are taken collectively and orders dispatched. Speech-recognition technology transcribes utterances to text and ASI is needed to label that text with speaker identities. An open-set ASI system is likely to be the best candidate for most C2 environments. Indeed the system would be trained to recognize the core regular staff. It would be tested for that core but also for infrequent visitors. The system would bunch the latter together under a single class 'other'. Ideally, the different C2 actors walk around within an environment free of microphone headsets or throat microphones. The wideband signal is captured by an unobtrusive omni-directional microphone placed at a fixed point inside the centre.

ASI SYSTEM OVERVIEW

Figure 1 gives an overview of an ASI system tailored for a C2 centre. The speech signal is reduced to vectors of parameters in the front end. Those vectors are classified as belonging or not to a speaker, in the back-end. The signal is

digitised and segmented. Silent as well as low-energy parts are automatically removed. (The signal may also, optionally, be high-frequency pre-emphasised to compensate for glottal and lip-radiation effects.) After a hamming windowing stage, vectors of cepstra are computed for each speaker.

The m^{th} cepstrum is given by:

$$c_m = \theta_c \sum_{j=0}^{J-1} \cos\left(m \frac{\pi}{J} (j+0.5)\right) \log_{10}(E_j) \quad (1)$$

where θ_c respects the dynamic range of the c_m , and E_j is the energy in each channel associated with the J triangular filters used.

The cepstrum is usually warped to a mel-scale (mel-cepstrum) [1]. The new parameter contains only static information. Vectors can be extended by the addition of dynamic information. This latter information is provided by the cepstrum's delta. Many formulas exist for computing the delta parameter. The standard m^{th} cepstrum's delta is given by [2]:

$$d_i = \frac{\sum_{t=1}^I i(c_{m,t+i} - c_{m,t-i})}{2 \sum_{i=1}^I i^2} \quad (2)$$

where $c_{m,t+i}$ is the static m^{th} cepstrum at time $t+i$, and i is the size of the window.

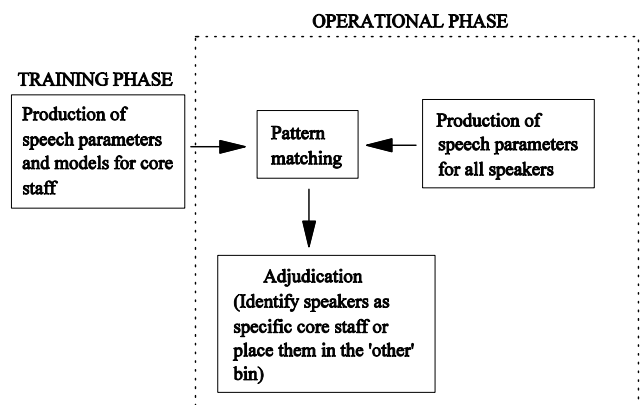


Figure 1. ASI system for C2 centre.

¹ Defence Science and Technology Organisation, PO Box 1500, Edinburgh, SA, 5111, Australia.