Institution:
University of Cambridge

Unit of Assessment:
UoA15

Title of case study:
Adaptation Techniques for Speech Recognition

1. Summary of the impact (indicative maximum 100 words)

Nearly every large-vocabulary speech recognition system in current use employs outputs from fundamental research carried out in the University of Cambridge Department of Engineering (DoEng) on adaptation of Hidden Markov Models (HMMs). One example of the commercial application of these outputs is their use on the Microsoft Windows desktop for both the command and control functions and the dictation functions. Approximately one billion copies of Windows have been shipped since 2008. Other examples show the outputs used in the automatic transcription of a wide range of types of data. [Text removed for publication]

2. Underpinning research (indicative maximum 500 words)

Phil Woodland started research on transform-based adaptation for speech recognition in 1993 at DoEng, having been appointed as a Lecturer in the DoEng in October 1992 (he started as an Assistant Lecturer in 1989 and was later promoted to Professor in 2002). This work led to a technique called Maximum Likelihood Linear Regression (MLLR) [1,2]. Mark Gales started working with Woodland in the DoEng as a Research Fellow in 1995. They worked on generalising MLLR [3]. Later in Gales' Fellowship, Gales developed transform-based adaptation and the Constrained MLLR (CMLLR) [4] technique. Gales left Cambridge in 1997 (after writing [4]) to work at IBM Research, but returned to the DoEng in 1999 as a Lecturer and was promoted to Professor in 2012.

Speech recognition systems have improved markedly over the last fifteen to twenty years, due to improvements in training techniques, the use of large amounts of training material, and improved computing resources. However, to obtain highly accurate models, it is important that a speech recognition system can quickly adapt the acoustic models it uses to better represent the characteristics of individual speakers and/or environmental conditions. This is particularly important if particular speakers/conditions are not well represented in the training data. The standard approach to speech recognition is based on the use of Hidden Markov Models (HMMs) to capture the variability of individual speech sounds in terms of a sequence of vectors that each represents the short-term spectrum and local time derivatives. Each of these vectors typically has a dimensionality of about forty. In a large vocabulary speech recognition system, there are a large number of HMMs that represent sounds in a particular phonetic context and can lead to hundreds of thousands of Gaussian components in the complete system. Normally, these Gaussians will have a diagonal covariance structure and, hence, the main parameters are Gaussian mean and variance vectors.

Speech recognition systems are conventionally trained using maximum likelihood (ML) estimation. The standard method of adaptation in the early 1990s used the maximum *a posteriori* (MAP) technique, which only adapts the Gaussians observed in the adaptation data and, hence, needs a relatively large amount of adaptation data to be effective. The original version of MLLR, developed in 1993-4 by Woodland's team, uses ML to estimate a set of full transform matrices and biases, which are applied to all the Gaussian means in the system and, hence, adapts Gaussian means not observed in the adaptation data. Even a speech recogniser with many millions of parameters can be effectively adapted with a few tens of seconds of adaptation data using MLLR. If more data is available, then more adaptation transforms can be reliably estimated using MLLR. Therefore, a variable number of transforms are used depending on the quantity of adaptation data and a flexible tree-based method of determining the number of transforms was developed [2]. MLLR was formulated in terms of extending the standard method of ML training of HMMs, which is an iterative approach updating the system parameters on each iteration. MLLR determines the statistics and performs a closed-form maximisation on each iteration to obtain a full maximum likelihood solution for MLLR using mean transforms for HMMs with separate variance vectors in each Gaussian component. The Gaussian mean parameters are the most important for adaptation: however, to accurately adapt models, especially to noisy audio, requires also adapting the variance parameters. Gales started working with Woodland in 1995 and extended the original version of MLLR to allow the Gaussian variance parameters to be also adapted [3] with a separate set of

transforms, using either a full variance transform or with diagonal transform. Furthermore, the mathematical analysis for MLLR method was extended to allow HMMs using full covariance matrices.

Gales continued working on transform-based adaptation. Constrained MLLR [4] estimates a consistent set of transforms that are applied to both the mean and the variance parameters. This allows the transforms to be applied to the acoustic features, and hence is sometimes referred to as feature MLLR, rather than to the model parameters themselves. This is a significant advantage for systems with large numbers of parameters and few transforms. It also means that it is straightforward to apply single transform adaptation in training since only the training feature stream needs to be altered. The development of CMLLR required extending the previous mathematical formulation used for MLLR and then the use of a novel iterative solution technique to finding the transform parameters.

Throughout the period described above, speech recognition research was greatly aided by annual evaluations organized by the US National Institute of Standards and Technology (NIST). These evaluations included entrants from research institutes, universities and companies. Each year the focus was on particular tasks: the transcription of read newspaper texts in the early 1990s; and later on the transcription of broadcast news (BN) data and conversational telephone speech (CTS). The Hidden Markov Model Toolkit (HTK)-based systems developed in Cambridge frequently had the lowest error rate on the main tests (in 1994, 1995 on newspaper dictation; in 1998, 2000, 2001, 2002 on CTS and in 1997, 2003 and 2004 on BN). The transform based adaptation methods described in this case study were used in the Cambridge HTK-based speech recognition systems developed for these evaluations from 1994 onwards. These included general adaptation to individual speakers (including non-native speakers) and different acoustic conditions.

The development of this family of techniques has continued at DoEng and led to a large number of improvements. These have included the use of lattice-based techniques for unsupervised adaptation [5] and discriminative estimation techniques [5,6] which can yield improved accuracy in some scenarios.

Research on transform-based adaptation has been carried out in the context of a number of research grants that have aimed to improve speech recognition technology. These have included those funded by EPSRC (1994-1997, 1997-2000); GCHQ (1996-2001); EU (2000-2003) and DARPA (2002-2007; 2005-2011). Woodland was the DoEng Principal Investigator for all of these grants. In addition to the above an EPSRC programme grant was awarded (2011-2016). The PI of this programme grant was from the University of Edinburgh, with Woodland as the lead Cambridge investigator.

3. References to the research (indicative maximum of six references)

1. *C. J. Leggetter and P. C. Woodland (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language, Vol 9, pp 171–185, DOI: 10.1006/csla.1995.0010. (Citations: 2217)
2. C. J. Leggetter and P. C. Woodland (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. Proc. ARPA Spoken Language Technology Workshop. pp 104-109. (Citations: 215)
3. *M.J.F. Gales and P.C. Woodland (1996). Mean and Variance Adaptation Within the MLLR Framework. Computer Speech and Language, Vol 10, pp 249–264, DOI: 10.1006/csla.1996.0013. (Citations: 403)
4. *M.J.F Gales (1998). Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition. Computer Speech and Language, Vol 12, pp 75-98, DOI: 10.1006/csla.1998.0043. (Citations: 944)
5. L.F. Uebel and P.C. Woodland (2001). Improvements in Linear Transform Based Speaker Adaptation. Proc. 2001 IEEE International Conference on Acoustics, Speech and Signal Processing. Vol 1, pp 49-52, DOI: 10.1109/ICASSP.2001.940764. (Citations: 58)
6. L. Wang and P.C. Woodland (2008). MPE-based Discriminative Linear Transforms for Speaker Adaptation. Computer Speech and Language, Vol 22, pp 256-272. (Citations: 23)

*Research outputs that best represent the quality of the research.

All citation counts are taken from Google Scholar. [1] is the most highly cited paper to have appeared in the journal Computer Speech and Language (CSL) and [3] is the second most highly

cited paper in CSL. When, in 2000, CSL introduced an annual award for the best paper published during the past 5 years it was awarded to [1].
Both Gales and Woodland are Fellows of the Institute of Electrical and Electronics Engineers (IEEE) and Woodland also became a Fellow of the International Speech Communication Association (ISCA). These honours are in part due to their work on transform-based adaptation. Woodland was invited to give plenary talks on speaker adaptation techniques at the following international workshops: 1999 IEEE International Workshop on Speech Recognition and Understanding, Keystone, Colorado, USA; and at the 2001 ISCA Workshop on Adaptation Methods for Speech Recognition, Sophia Antipolis, France.

4. Details of the impact (indicative maximum 750 words)

The papers on MLLR have been highly influential in both a research and commercial context: the methods have become part of the standard approach to speech recognition and used by most systems that perform any type of adaptation. The techniques are covered in standard textbooks, e.g. [7], and courses given on speech recognition, e.g. [8,9].
There are a number of different scenarios that describe how adaptation can be applied in a speech recognition system. If the word-level transcription of the adaptation data is known, then it is termed supervised adaptation, and, if it has to be estimated by a recognition pass, this is unsupervised. Transcription systems that do not require very low latency output typically use multiple passes through the data, with an initial recognition pass using un-adapted models, which gives the transcription used for estimating adaptation transforms for a later pass. In this case, it is essential that the adaptation is robust to errors in the first pass transcription, and that it is effective with small amounts of adaptation. MLLR and CMLLR are widely used for this purpose. In some applications that include supervised adaptation at enrolment time, the adaptation information can be further updated in an incremental fashion using unsupervised adaptation, for example, to update the speaker profile associated with a particular speaker.
Throughout the research, the DoEng speech group has developed versions of the HTK. This has been available for free download since September 2000 (from http://htk.eng.cam.ac.uk). Version 3.3, released in 2003, incorporated all the research referenced in [1,2,3,4]. Version 3.4.1 is the latest version released in 2009. The software is used by some companies to develop new products and services as well as in more fundamental research in a variety of research labs around the world. The software currently has over 100,000 licensees.
Many companies have used MLLR and CMLLR adaptation in speech recognition systems for a wide range of use cases. Some examples are described below, although note that some companies have asked for their support statements to remain confidential.
[Text removed for publication] [10]
Speech recognition can be used as part of the interface for desktop and laptop computers. Microsoft introduced a new speech interface, Windows Speech Recognition for Vista, as part of the Windows Vista operating system launched in 2007. The speech recognition engine is available in eight languages. This provides both command and control of Windows functions by voice as well as the ability to dictate text and it is deeply integrated into the operating system. In order to be effective, it is vital that the system has high accuracy. There are two main phases of acoustic model adaptation: speaker enrolment is used and also unsupervised adaptation is applied during normal use. The initial adaptation requires a script to be read out loud by the user. Later adaptation includes feedback from corrections and alternate selections to refine the adaptation process. In all cases, MLLR is used in the acoustic model adaptation process (along with maximum a posteriori adaptation). The same speech recognition technology is also an integral part of Windows 7 and Windows 8. Total sales for Windows between January 2008 and July 2013 are approximately one billion licences which include 630 million copies of Windows 7 between its launch in October 2009 and July 2012 [11].
[Text removed for publication] [10,12,13]

5. Sources to corroborate the impact (indicative maximum of 10 references)

7. D. Jurafsky & J.H. Martin, Speech and Language Processing, Prentice Hall, 2nd Edition, ISBN978-0131873216, 2008
8. Speech Recognition and Synthesis Winter 2009, Course Information, Stanford University website, http://www.stanford.edu/class/cs224s/
9. Automatic Speech Recognition (ASR): 2012/13, Course Descriptor, Edinburgh University

website, http://www.inf.ed.ac.uk/teaching/courses/asr

10. Statement by Chief Technology Officer at Nuance Communications.
11. Statement by an Architect in the Microsoft speech team and the Partner Engineering Manager leading the Microsoft speech product team.
12. Statement by Manager of Speech Processing Research at the IBM Thomas J. Watson Research Center.
13. Statement by Chief Scientist, Raytheon BBN Technologies.