Evaluation et avancées en reconnaissance de la parole: de "Resource Management" à "Broadcast News"

J.L. Gauvain

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
gauvain@limsi.fr

RÉSUMÉ

Le programme d'évaluation le plus important sur la reconnaissance de la parole est organisé par ARPA (Advance Research Project Agency) aux Etats-Unis, programme dans lequel le NIST (National Institute for Science and Technology) coordonne le déroulement des tests. Le paradigme d'évaluation mis en œuvre par ARPA a pour but d'accélérer les développements technologiques, en offrant à la communauté scientifique les corpus nécessaires au développement des modèles et systèmes ainsi qu'une infrastucture pour mesurer et comparer les performances de ces différents systèmes.

Dans cet exposé, sont examinés les progrès réalisés grâce à ces campagnes d'évaluations, en particulier à travers l'expérience du LIMSI qui y a pris part pour la première fois en 1992.

Les premières évaluations (de 1987 à 1992) ont été effectuées sur la tâche "Resource Management" (RM) avec un vocabulaire de 1000 mots [7], puis pour des vocabulaires de 5000 et 20000 mots sur le corpus "Wall Street Journal" (WSJ) [8, 9], et plus récemment pour un vocabulaire illimité sur la tâche "North American Business News" (NAB) [10, 11]. En 1995 les données de test ont été enregistrées avec 8 microphones différents dans un environnement acoustique bruité (corpus MUM) [11] afin d'inciter les développements rendant les systèmes plus robustes aux changements de microphone et d'environnement acoustique.

Pour les évaluations RM, WSJ-5k, et WSJ-20k, un certain nombre de contraintes ont été imposées pour les tests primaires ("baseline tests"): contraintes sur les données d'apprentissage, le vocabulaire ou les modèles de langage, dans le but de focaliser les développements sur le problème spécifique de la modélisation acoustique. Par la suite ces contraintes ont été progressivement retirées pour finalement laisser aux chercheurs et développeurs de systèmes toutes libertés pour mettre en œuvre les solutions les plus performantes.

Grâce à ces campagnes d'évaluations, des progrès significatifs ont ainsi pu être mesurés sur le problème de la reconnaissance mutilocuteur de la parole continue. La progression mesurée en terme de réduction du taux d'erreur est particulièrement évidente entre 1987 et 1992, années pendant lesquelles toutes les évaluations ont été effectuées sur la même tâche (RM). Le taux d'erreur sur cette tâche est passé en quelques années de plus de 20% à 4%. Par la suite on a vu la tâche proposée par ARPA se complexifier mais aussi devenir plus réaliste faisant ainsi faire à la communauté scientifique des progrès importants en matière d'architecture des décodeurs (aujourd'hui capables de traiter des vocabulaire de 65k mots) et de techniques d'adaptation des modèles acoustiques.

En 1995, ARPA a proposé la tâche "Broadcast News" (BN) pour la transcription d'émissions radio et télédiffusées. Cette tâche se démarque des précédentes par l'utilisation de données qui ne sont pas artificielles c'est-à-dire produites aux seules fins de l'évaluation. (La quasi-totalité des données d'évaluation pour les tâches RM, WSJ, et NAB était constituée de parole "lue".) Pour préparer l'évaluation de novembre 1996, le LDC (Linguistic Data Consortium) a mis à disposition des participants environ 100 heures d'émissions dont 50% étaient transcrites orthographiquement. Ces données contiennent des segments de différentes natures: parole "lue", parole préparée, parole spontanée, enregistrements de qualité studio, enregistrement téléphoniques, parole sur fond musical, musique, etc...

Les résultats des six dernières évaluations organisées par ARPA sont donnés dans le tableau 1. Le nombre de participants chaque année varie entre 6 et 12. Les sites ayant participés au moins deux fois sont: AT&T, BBN, BU, CMU, CUED htk/con, Dragon, IBM, LIMSI, Lincoln lab, NYU, Philips, Rutgers, SRI. Le LIMSI a participé à ces six dernières évaluations [3, 5, 4, 1].

Les résultats présentés ici correspondent seulement aux conditions primaires qui sont retenues pour comparer les différents systèmes. Pour chaque évaluation plusieurs conditions contrastives ont également été testées (comparant par exemple systèmes mono et multilocuteur pour le corpus RM). Sur ce point, on peut noter en particulier le paradigme "Hub and Spokes" de l'évaluation de novembre 1994 où de nombreuses conditions contrastives ont été testées: parole lue *vs* parole spontanée, large bande *vs* téléphone, Sennheiser *vs* microphone inconnu, et différents rapports signal sur bruit [10].

Test	*Conditions*	*Vocabulaire*	*Taux erreur(%)*
Sep92 RM	1k "wordpair", vocabulaire fermé	1k	4.4 - 11.7
Nov92 WSJ	5k bg, vocabulaire ouvert	5k	6.9 - 15.0
	20k bg	20k	15.2 - 25.2
Nov93 WSJ	20k open tg	20k	11.7 - 19.0
	5k bg	5k	8.7 - 17.7
	5k tg	5k	4.9 - 9.2
	5k tg, téléphone local	5k	12.8 - 25.5
Nov94 NAB	20k tg, illimité	20k	10.5 - 22.8
	illimité	20 - 65k	7.2 - 17.4
	illimité, téléphone	40 - 65k	22.5 - 24.6
Nov95 MUM	illimité, bruit, mic inconnu	65k	13.5 - 55.5
	illimité, bruit, Sennheiser	65k	6.6 - 20.2
Nov96 BN	illimité, bruit, mic inconnu	65k	27.1 - 53.8
	parole spontanée, musique, téléphone		

Table 1: Résultats (taux d'erreur de mot) des évaluations ARPA entre 1992 et 1996 pour les conditions primaires. Les tests ont été effectués sur des tâches de difficulté croissante. (Les conditions "bg" et "tg" correspondent à des modèles de langage bigramme et trigramme.) Le taux d'erreur le plus faible et le plus elevé sont donnés pour chaque test.[1]

Plusieurs remarques doivent être faites à propos de ces résultats. Premièrement, on peut noter que pour les tests à vocabulaire fermé les taux d'erreur sont particulièrement bas, 4% pour un vocabulaire de 1000 mots et 6% pour 5000 mots. Deuxièment, l'augmentation de la taille du vocabulaire ne réduit pas les performances (à condition que le modèle de langage soit proprement construit), ceci a été clairement démontré lors de l'évaluation de 1994. Troisièmement, alors que ce tableau de résultats ne contient que les taux d'erreur moyens, on observe communément pour un même système un rapport 10 entre les taux d'erreur pour le meilleur et le plus mauvais locuteur. Enfin alors que les tests ont été élargis à des conditions proches d'applications réelles (téléphone, parole bruitée, microphone inconnu, dictée spontanée), des progrès substantiels sont encore nécessaires pour utiliser ces systèmes dans le monde réel [2].

Les résultats obtenus sur la tâche BN montrent que l'on peut aujourd'hui transcrire (sans contrainte "temps réel") des émissions d'information en langue anglaise avec un taux d'erreur de l'ordre de 30%. Ce résultat bien qu'insuffisant pour envisager de générer automatiquement des transcriptions exactes nous laisse cependant entrevoir la possibilité à moyen terme d'indexation automatique de documents audiovisuels [6].

Une des principales sources de progrès est bien entendu la disponibilité de très grands corpus de parole à partir desquels il nous est possible d'estimer les paramètres de modèles acoustiques et de modèles de langages de plus en plus complexes (typiquement quelques millions de paramètres).

RÉFÉRENCES

[1] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News Shows", *Proc. IEEE ICASSP-97*, Munich, Avr. 1997.

[2] J.L. Gauvain, L.F. Lamel, "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications", *IEICE*, Dec. 1996.

[3] J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), Oct. 1994.

[4] J.L. Gauvain, L.F. Lamel, G. Adda and D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *Proc. ICASSP-96*, Atlanta, GA, Mai 1996.

[5] J.L. Gauvain, L.F. Lamel and M. Adda-Decker, "Developments in Large Vocabulary Dictation: The LIMSI Nov94 NAB System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Jan. 1995.

[6] A. Hauptmann, and H. Watlar, "Indexing and search of multimodal information," *IEEE ICASSP-97*. Munich, Avr. 1997.

[7] D.S. Pallett, J.G. Fiscus and J.S. Garofolo, "Resource Management Corpus: September 1992 Test Set Benchmark Results," *Proc. ARPA Workshop on Continuous Speech Recognition*, Stanford, CA, Sep. 1992.

[8] D.S. Pallett, J.G. Fiscus, W.M. Fisher, and J.S. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, Mars 1993.

[9] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, and M.A. Pryzbocki, "1993 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, Mars 1994.

[10] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, A.F. Martin and M.A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Jan. 1995.

[11] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, A.F. Martin and M.A. Przybocki, "1995 Hub-3 Multiple Microphone Corpus Benchmark Tests," *Proc. ARPA Speech Recognition Workshop*, Harriman, NY, Fév. 1996.

[1] ARPA organise également des évaluations sur la parole conversationelle en utilisant le corpus Switchboard. Les taux d'erreur sont actuellement de l'ordre de 45% sur l'anglais américain.