Speech Processing et SEO : l'avenir du référencement audio (2ème partie)

Par Yann Sauvageon

Domaine :	Recherche	**Référencement**
Niveau :	**Pour tous**	Avancé

Le référencement audio est l'un des grands sujets du SEO dans les années qui viennent, avec l'omniprésence de plus en plus forte de la vidéo dans notre vie numérique. Au programme de cet article en deux parties : chiffres clés sur la vidéo, historique du Speech Processing chez Google, les dernières avancées en date et notamment le Deep Learning, ainsi, bien sûr, que les applications opérationnelles en SEO aujourd'hui....

Le mois dernier, nous avons décrit l'« état de l'art » du référencement audio et du *speech processing*. Il est maintenant temps de se pencher sur l'avenir et notamment les étonnantes possibilités du *deep learning* en la matière...

De nouvelles perspectives grâce au Deep Learning...

Les dernières évolutions technologiques et notamment le « Deep Learning » laissent entrevoir pour le Speech Processing une amélioration probablement plus rapide des performances que ce qui était envisagé jusqu'à présent.

Andrew NG, le génie du Deep Learning, à l'origine de Google Brain et désormais Directeur scientifique chez Baidu, ne s'y trompe pas. Pour lui, le Speech Processing est l'un des champs d'application majeur du Deep Learning.

Le Deep Learning ? Kesako ? Nous ne prétendons pas être un expert du sujet mais lançons-nous quand même dans une définition. Le Deep Learning repose sur des réseaux neuronaux (artificiels), organisés en plusieurs couches de neurones effectuant un apprentissage hiérarchique des caractéristiques.

La grande nouveauté, par rapport aux techniques classiques de Machine Learning, est que les algorithmes de Deep Learning apprennent seuls et sont capables d'extraire les caractéristiques d'un input (un son, une image, un texte, ...) sans qu'il y ait besoin de fournir des données de départ taguées manuellement.

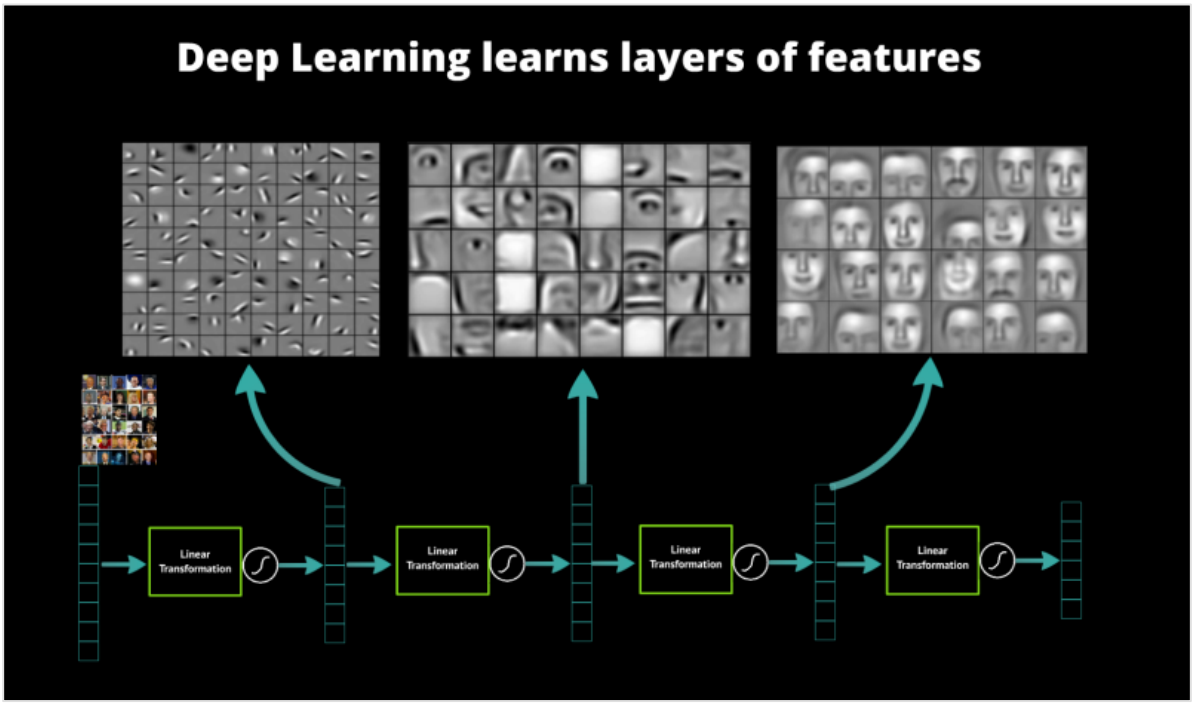
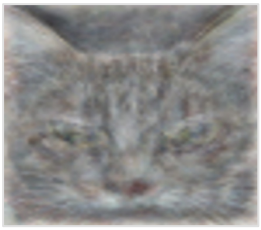


Fig. 1. Illustration de l'apprentissage hiérarchique
(source : http://www.datarobot.com/blog/a-primer-on-deep-learning/)

Le Deep Learning n'est pas qu'une tendance à la mode, mais bien une technique explorée par les plus grands : Cortana (Microsoft : http://www.01net.com/editorial/627374/bernard-ourghanlian-demain-un-assistant-personnel-pourra-agir-a-notre-place/), Facebook (http://www.atelier.net/trends/articles/facebook-adopte-deep-learning-mieux-comprendre-utilisateurs_424287), Skype, Baidu (http://finance.yahoo.com/news/baidu-says-massive-deep-learning-231743161.html) ou Nvidia (http://www.forbes.com/sites/patrickmoorhead/2015/03/24/nvidia-bets-big-on-deep-learning/)…

Un des faits d'arme les plus connus reste l'expérience montée par Google en 2012 qui a fait tourner un réseau neuronal de 16 000 machines. La mission de ce réseau : regarder des vidéos YouTube non-stop pendant 1 semaine. A l'issue de cette semaine, Google a découvert qu'un des neurones artificiels avait découvert ce qu'était un chat, sans qu'on lui ait dit au préalable ce qu'était cet animal ou les caractéristiques constitutives d'un chat ! (http://googleblog.blogspot.fr/2012/06/using-large-scale-brain-simulations-for.html) C'est à l'époque le fameux Andrew NG qui menait les opérations.



Les promesses du Deep Learning relatives au Speech Processing sont énormes

Déjà, en 2010, Microsoft avait annoncé des améliorations significatives de leur speech recognition en utilisant le Deep Learning comme le témoigne la figure 2.

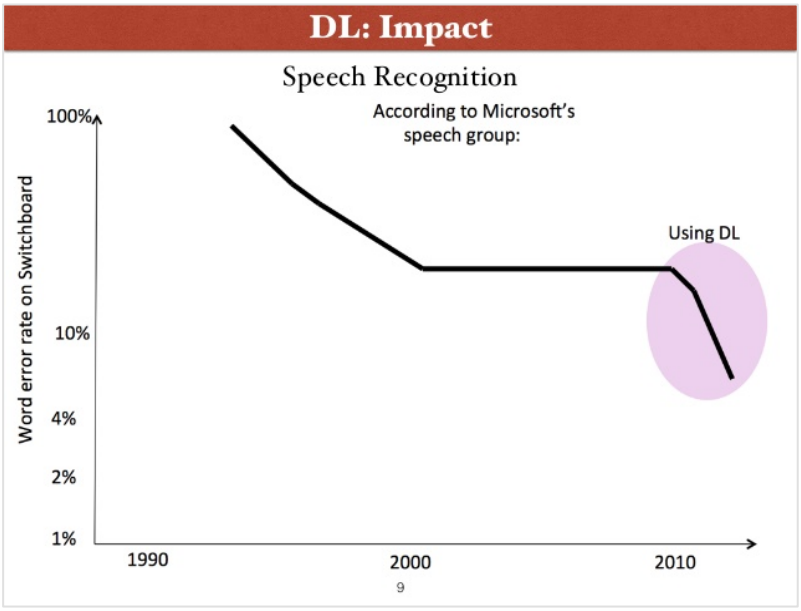


Fig. 2. Impact du Deep Learing en speech recognnition (source : http://fr.slideshare.net/roelofp/220115dlmeetup)

Aujourd'hui, Baidu investit largement dans cette voie avec une équipe dédiée, managée par Andrew NG, et des systèmes de Deep Learning 100 fois plus puissant que ceux déployés par Google en 2012.

Avec son système de Speech Recognition « Deep Speech », Baidu annonce début 2015 des performances 10% supérieures à celles de ses concurrents (Google Speech API, Bing Speech, Apple Dictation …) notamment grâce au Deep Learning (fig.3).

Guerre des chiffres et effet de communication, la réponse de Google ne saurait tarder.

Impacts SEO

Revenons à nos moutons et aux enjeux concrets au niveau SEO. Aujourd'hui, les moteurs de recherche ne crawlent pas l'audio des vidéos et ne proposent pas un système d'indexation in-video. En conséquence, pour comprendre une vidéo, les moteurs de recherche continuent de se baser sur les éléments on-page concomitants au fichier (Title, Description, Texte à proximité de la vidéo, transcript écrit associé à la vidéo).

Quant à YouTube, qui fournit un transcript automatique, Brad Ellis nous confirmait fin 2013 que Google n'utilisait pas les transcripts automatiques pour indexer vos contenus vidéos, du fait de taux d'erreur encore trop importants.

Ainsi, la conclusion opérationnelle est qu'à court terme, on ne peut pas compter sur les moteurs de recherche pour indexer nos vidéos et qu'il va falloir continuer de nourrir les moteurs avec du texte connexe.

System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**

Table 4: Results (%WER) for 3 systems evaluated on the original audio. All systems are scored *only* on utterances with predictions given by all systems. The number in parentheses next to each dataset, e.g. Clean (94), is the number of utterances scored.

Fig. 3.

Source : https://gigaom.com/2014/12/18/baidu-claims-deep-learning-breakthrough-with-deep-speech/)

Deux enjeux SEO majeurs se profilent :
• Fournir un transcript fiable à 100% ;
• La catégorisation de ses vidéos (lorsque le volume est important).

Fournir un transcript fiable à 100%

Le premier enjeu est clairement de fournir un transcript fiable à 100% qui pourra être associé à votre page (comme on peut le voir sur les vidéos TED par exemple, fig.4) ou qui pourra être uploadé sur YouTube (fig.5).

Pour fournir un transcript fiable, la meilleure méthode est de combiner un transcript automatique avec une correction humaine en sortie. Les budgets oscillent entre 0.80 € et 2.50 € la minute et peuvent intégrer uniquement le transcript automatique ou le package transcript + correction manuelle.

A noter qu'il peut être intéressant d'anticiper un éventuel crawl futur des transcripts par Google et consorts en adoptant dès à présent le balisage HTML5 « Track » (fig.6) qui permet d'associer à une vidéo un fichier de transcript (sous-titres ou captions) en utilisant le format WebVTT (http://dev.w3.org/html5/webvtt/). Au sein de ce fichier, vous pourrez renseigner le transcript texte de votre vidéo mais également préciser des indications temporelles permettant de créer des ancres pointant directement vers des sections au sein de la vidéo.

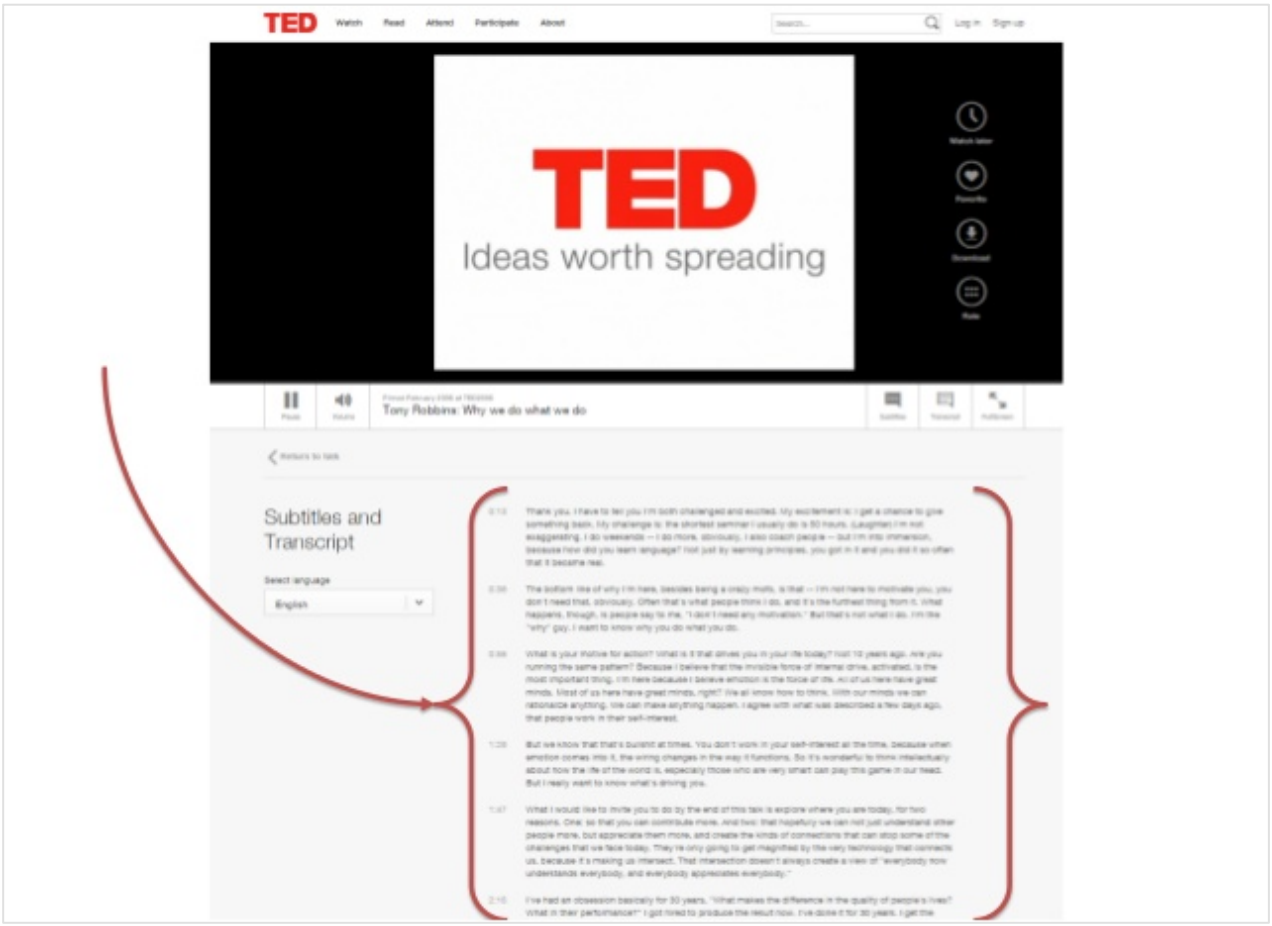


Fig. 4. Exemple d'un transcript texte sur le site TED.Com.

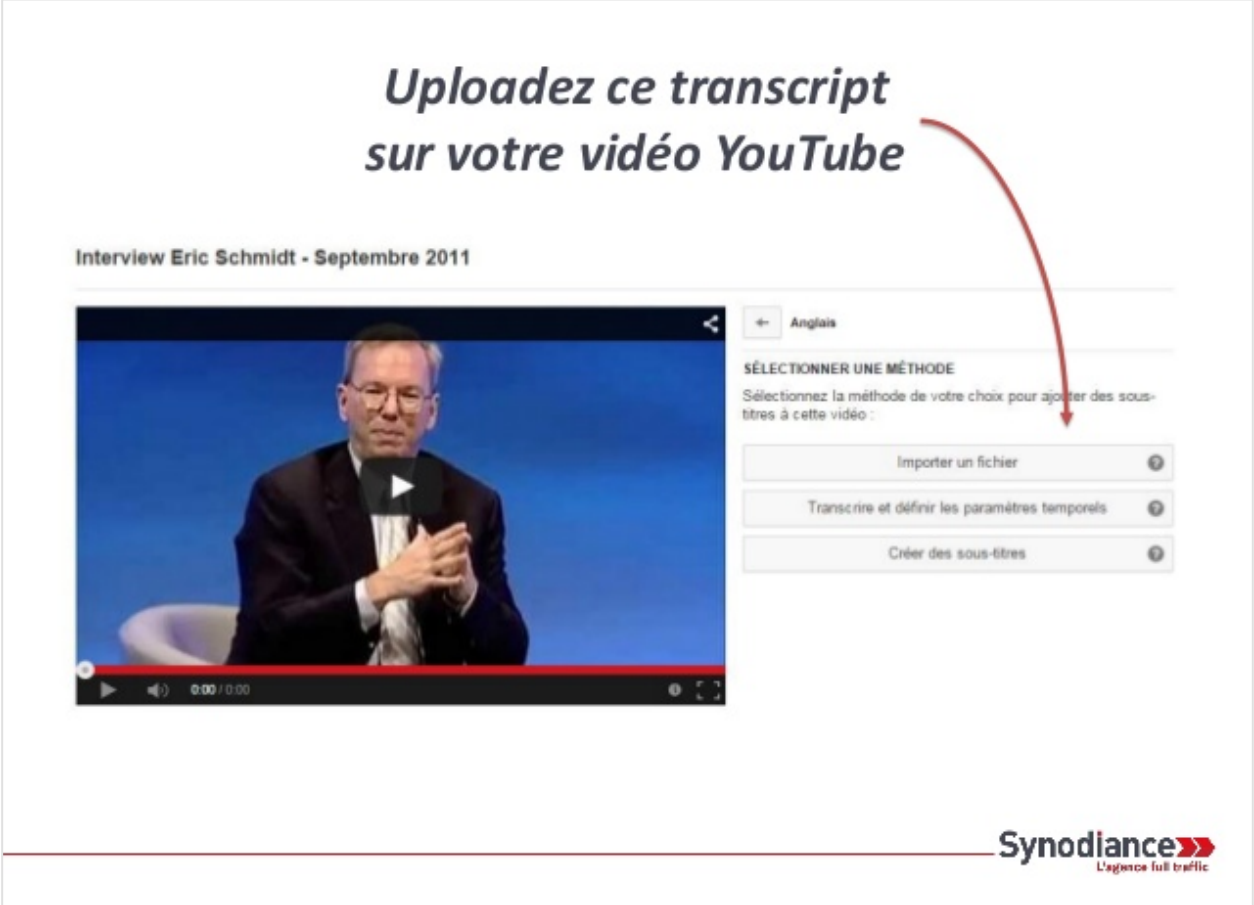


Fig. 5. Le transcript texte peut être uploadé directement sur YouTube.

```
<video id="video" controls preload="metadata">

  <source src="video/sintel-short.mp4" type="video/mp4">
  <source src="video/sintel-short.webm" type="video/webm">

  <track label="English" kind="captions" srclang="en" src="captions/vtt/sintel-en.vtt" default>
  <track label="Deutsch" kind="captions" srclang="de" src="captions/vtt/sintel-de.vtt">
  <track label="Español" kind="captions" srclang="es" src="captions/vtt/sintel-es.vtt">

</video>
```

Fig. 6. La balise <track> en HTML5.

La catégorisation SEO de ses vidéos

Cet enjeu s'applique aux acteurs bénéficiant d'un nombre élevé de vidéos. L'enjeu ici n'est pas de réaliser un transcript audio fiable à 100% mais d'utiliser les technologies de Speech Processing pour extraire les thématiques saillantes ou les mots clés principaux de vos vidéos.

Vous bénéficiez de milliers de vidéos non taguées ou partiellement taguées. Vous souhaitez extraire de ces vidéos celles parlant d'un sujet spécifique ou intégrant un mot clé spécifique. Le Speech processing vous permettra de réaliser ce travail de ma-

nière extrêmement rapide et à un coût compétitif.

Exemple : Vous avez des vidéos de recettes de cuisine et souhaitez sous-catégoriser vos vidéos par épice (aneth, basilic, cannelle, ciboulette, citronnelle, coriandre, gingembre, romarin, thym…). Le Speech Processing va vous permettre d'identifier rapidement les vidéos citant ces épices et ainsi de les taguer automatiquement. En sortie, grâce à ce taggage automatisé, il est possible de créer des pages hub « éditorialisées » pour chaque épice et ainsi multiplier ses opportunités de positionnement.

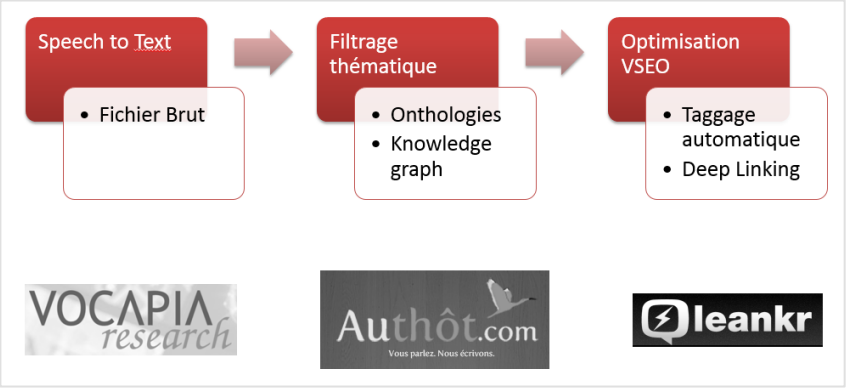


Fig. 7. Quelques acteurs du Speech Processing en France.

Autre exemple : Ceci peut s'avérer également très intéressant pour des sites d'actualité avec la possibilité de rechercher au sein des vidéos les sujets qui font l'actualité et ainsi sous-catégoriser à la volée en fonction de l'actualité.

Les acteurs

Il existe quelques acteurs français spécialisés dans tout ou partie de ce process de speech processing.
A titre informatif :

- **Vocapia** (http://www.vocapia.com/) va par exemple se concentrer sur la livraison d'un transcript texte brut.
- **Authot** (http://www.authot.com/fr/) va apporter en plus une interface de correction manuelle permettant un traitement plus rapide du transcript.
- **Leankr** (http://leankr.com/) va se connecter avec des knowledge graph et apporter un enrichissement au flux audio *via* des bulles informatives.

Conclusion

Le Speech processing devrait connaitre dans les mois à venir des avancées significatives. Est-ce que cela se traduira par un crawl des contenus vidéo par Google, l'avenir nous le dira. En tout cas, dans un monde de plus en plus vidéo, le SEO doit trouver les solutions adaptées et anticiper les changements à venir :

- Sous-traiter ses transcripts en faisant appel au speech processing ;
- Sous catégoriser ses vidéos en extrayant rapidement les entités clés de vos vidéos ;
- Baliser d'ores et déjà vos fichiers de sous-titres via la balise track.

Ce sont là quelques-unes des pistes à travailler.

Yann Sauvageon, *Directeur de l'expertise, Synodiance (http://www.synodiance.com/).*

Cet article fait suite à la conférence co-animée par l'auteur avec Jérôme Rocheteau au SEO Campus 2015, dont vous trouverez les slides ici :
http://www.slideshare.net/Synodiance/synodiance-seo-et-speech-processing-futur-enjeu-seo-seo-campus-2015